5-2021

Genomics-Assisted Breeding for Grain Yield and Composition in Sorghum

Sirjan Kumar Sapkota
Clemson University, sirjan.sapkota@yahoo.com

Genomics-Assisted Breeding for Grain Yield and Composition in Sorghum

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Plant and Environmental Sciences

---

by
Sirjan Kumar Sapkota
May 2021

Accepted by:
Stephen Kresovich, PhD, Committee Chair
Richard E. Boyles, PhD
William Bridges, PhD
Benjamin Todd Campbell, PhD
Ksenija Gasic, PhD

Abstract

Cereal grains provide over half of the total calories for human and animal nutrition. Sorghum [*Sorghum bicolor* (L.) Moench] is the fifth most important cereal grain in the world and a source of staple for over half a billion people in the semi-arid tropics. As human population is projected to become nine billion by middle of this century, crop production needs to increase by 70% to 100% to meet the increasing demand for food. The advancement in genomic technologies and their application in breeding has potential to assure food security. The objectives of this study was to explore application of whole genome markers in identifying marker trait associations, potential gene candidates associated with the traits, and evaluating prediction performance of whole genome regression models in sorghum. Grain yield and grain composition traits measured in multiple environments and populations were used in model training and cross-validation of prediction performance using different statistical approaches. In general, genomic prediction for grain yield components and grain composition showed moderate to high accuracy depending on trait genetic architecture. Prediction accuracy of yield components declined when population structure was controlled. Race explained upto 50% of covariance for grain and panicle traits, and subpopulation with high genetic diversity had higher prediction accuracy. The prediction accuracy of grain composition for multi-trait model increased by 30-40% on average over single-trait model, suggesting multi-trait models using traits strongly correlated can increase genetic gain. A novel genomic association for starch was identified ~52 Mb of chromosome 8, and five out of six associated variants were located within a heat shock protein 90, *Sobic.008G111600*. Multivariate association for starch and protein identified additional variants around 60 Mb of chromosome 4, including one within 5'UTR of a fatty acid desaturase gene, *Sobic.004G260800*. Our results show genomic prediction can improve accuracy of selection in sorghum breeding and multivariate analysis of correlated traits can benefit association and prediction models.

Dedication

To Kathleen, for all the years of love and support you have bestowed upon me. <3 :× <3

Acknowledgments

First of all, I would like to thank and congratulate all my teachers and mentors for inculcating the culture of learning and curiosity of knowledge. The most immediate of that group of people were my graduate committee members who have pushed me to think hard and holistically in the matter of scientific process, philosophy and impacts. Specially Rick, who mentored me on and off the field through out my doctoral research, and I hope to reciprocate by providing mentorship to any student who might seek my guidance. When I started my doctoral degree I lacked the understanding of bigger picture of scientific process and philosophy, so I was obviously a little distraught when I had to start thinking independently about research questions and formulate hypothesis for my doctoral research. Steve pushed me hard to learn and see what is that I lacked and where I desired to be. His advisement and mentorship has made me more inquisitive and critical because I now understand that science is never complete but is always sufficient.

Second group of people I cannot thank enough are my friends and co-workers who has provided me the environment where I can grow. All the members of the Kresovich lab were monumental in my success in obtaining this degree, so I would like to extend my heartfelt gratitude to all of them. My friends have been the buffer that keep me burning from sometimes caustic environment of scientific process, I would like to thank them all enormously.

I have been lucky to have a very loving and supportive family, and I am very happy to be their pride. They keep me humble and grounded, and give me unconditional support which I have always been grateful for and will continue to be thankful by reciprocating my love and support.

I saved the best for the last because I have always tried to delay gratification since I read Dale Carnegie's book in college. My wife, Kathleen, who is by far the most invested person in my personal and professional life after me. We spent five years separated by about 300 miles but lived our lives like one heart and one soul. She has given me the courage to face my fears and grow with

love towards everything. This is not a hyperbole, I would have never been able to finish my degree if I didn't have a partner who was so kind, loving and supportive. I love you Kathleen and will always love you.

Table of Contents

List of Tables

Page

List of Figures

Page

Chapter 1

Introduction

1.1 Background

This century is going to be defined by two important paradigms, one opportunity and the other a challenge. Both are unprecedented in the history of our species and were brought about by our species. The opportunity is that we are technologically the most advanced civilization that we know of, and as a species we are capable of overcoming the direst of circumstances. The challenge is, sadly, us as an entity of nature and our impact in our environment in the light of our own survival. As a crop scientist, I am going to illuminate on these paradigms from the perspective of crop improvement.

Agriculture has been the foundation of our civilization. The domestication of crops led to the surplus of food and changed life style and culture of our species. In the last four decades of the twentieth century, the total food production was doubled despite decline in total farming population and farmed land (Khush, 2001). While several factors such as agronomic practices and assess to fertilizers and pesticides played a significant role in these, genetic improvement was a major contributor to increased crop yield potential and stability (Khush, 2001). In my undergraduate class, 'Introduction to Plant Breeding', we defined plant breeding as an art and science, and it is probably still considered to be so by many plant breeders. However, a large number of plant breeders will also agree that, with increased role of genetics and statistics, plant breeding today has definitely shifted heavily towards being more of a science than art. Furthermore, the accessibility to whole genome marker data due to the decline in sequencing cost and powerful statistical learning tools in modeling genetic data has resulted in increasingly data-driven plant breeding methodology.

Human population has quadrupled in less than a century, and a rapid improvement in agricultural practices, technology and germplasm was required during the mid-twentieth century to feed the growing population (Khush, 2001). As population grows to a projected 9 billion by 2050, we will, yet again, need to double our current food production to meet the demand (Godfray et al., 2010). While increasing the yield potential is already difficult, it is going to be even more daunting as climate change is rendering the planet hotter and drier.

Here in my doctoral dissertation, I am going to discuss a crop that is an important staple for hunger-striven semi-arid tropics and plant breeding tools that have started a new paradigm in plant and animal breeding. The opportunities this crop and these genomic tools provide will be crucial in efforts to overcome the challenges of hunger in this century.

1.2 Sorghum

1.2.1 Origin, domestication, and evolution

Sorghum, a genus in the grass family poaceae, belongs to the Andropogoneae tribe which is popular for its C_4 photosynthesis (Paterson et al., 2009). Ancestral sorghum genome diverged from the ancestor of rice genome around 50 million years ago (mya) (Wolfe et al., 1989) and the ancestral maize genome about 11.9 mya (Swigoňová et al., 2004). The genus *Sorghum*, after splitting from its progenitors, has undergone several domestication and diversification events giving rise to several species within the genus. However, the term 'sorghum' is commonly used to refer to the cereal crop species *Sorghum bicolor* (L.) [Moench]. The oldest record of cultivated sorghum is the charred remains of sorghum seeds discovered during archaeological excavations at Nabta Playa site near the Egyptian–Sudanese border (Wendorf et al., 1992). The seeds were dated to be from about 8,000 years before present (bp) and consisted of several wild as well as domesticated races of sorghum (Dahlberg et al., 1996). Based on the archaeological evidence and genetic diversity, the Sahel region of sub-Saharan Africa is considered to be the region where early domestication of cultivated sorghum occurred (Kimber et al., 2013). However, after domestication, further migration and adaptation of early sorghum domesticates continued across Africa and Asia (Figure 1.1A). The evolution of morphologically and geographically diverse groups, that are classified into five major races and 10 intermediate races, are hypothesized to be the result of evolutionary diversification of early sorghum domesticates due to such demographic events (Harlan et al., 1976; Harlan et al., 1972). The panicle and grains of sorghum belonging to different races can vary widely in shape and sizes, and are commonly used as means for racial classification (Figure 1.1B) (Harlan et al., 1976; Kimber et al., 2013). The five major races of sorghum are bicolor, caudatum, durra, guinea, and kafir. Among these races, bicolor is thought to be the most primitive and possibly the earliest domesticate that diversified into various racial types because of its widespread distribution across all sorghum growing regions of the world (Kimber et al., 2013). The evolution and diversification of panicle architecture among different sorghum races is an example of adaptation of cultivated sorghum (Morris et al., 2013). For example, guinea races that are adapted to west Africa, with high rainfall and humid climate, have wider panicles with pendulous and open branches, whereas, durra sorghum that evolved in arid regions of southern India has more compact panicles. The distinct demographic history and morphological diversification of sorghum has led to different crop ideotype

for different sorghum growing regions.

1.2.2 Cultivation and importance

Sorghum has been cultivated in the African continent since domestication of the crop and in Asia since its introduction through migration. Sorghum is a drought-tolerant crop requiring little input during growth but has the potential to yield better with good husbandry. The first sorghums introduced to the United States (US) were carried across the Atlantic with the slave-trade and reached US via the Caribbean islands (Doggett, 1988). By the late nineteenth century African landraces were grown in the US by the names of Milo maize, Guinea kafir, and Gyp corn (white durras). Because these cultivars were too tall and late maturing to be of much use, farmers selected and multiplied lines with mutations affecting height and maturity (see more in Doggett, 1988). The early African landrace introductions leading to grain sorghum development included Blackhull Kafir (1890), Feterita (1906), Giant Milo (1879), Hegari (1908), Pink Kafir (1904), White and Brown Durra (1874), and White and Red Kafir (1876) (Maunder, 2000). Today, sorghum is predominantly grown in the high plains of west Texas, Oklahoma, Kansas and Nebraska. Among all countries, US is the largest producer of sorghum grain (9.28 million tons) followed by Nigeria (6.86 million tons), however, Nigeria (6 million hectares) had three times the area harvested compared to US (FAOSTAT, 2018).

Sorghum is traditionally grown as a crop with diverse end-uses such as food, feed, fiber, fuel, and forage (Doggett, 1988). In vulnerable regions of semi-arid tropics of Africa and South Asia, sorghum is a staple source of nutrition for over half a billion people (Mace et al., 2013). While most of the sorghum grown in Africa and South Asia is consumed as a food staple, the sorghum produced in industrialized economies such as US and Australia is predominantly used as animal feed for the livestock industry. Nonetheless, as a starch-rich gluten-free grain with nutraceutical properties, sorghum is gaining popularity among food and beverage industries as a speciality crop (Taylor et al., 2006; Zhu, 2014). Sorghum has also gained popularity as a bioenergy crop due to its potential for high biomass accumulation and sugar retention in stalks (Brenton et al., 2016).

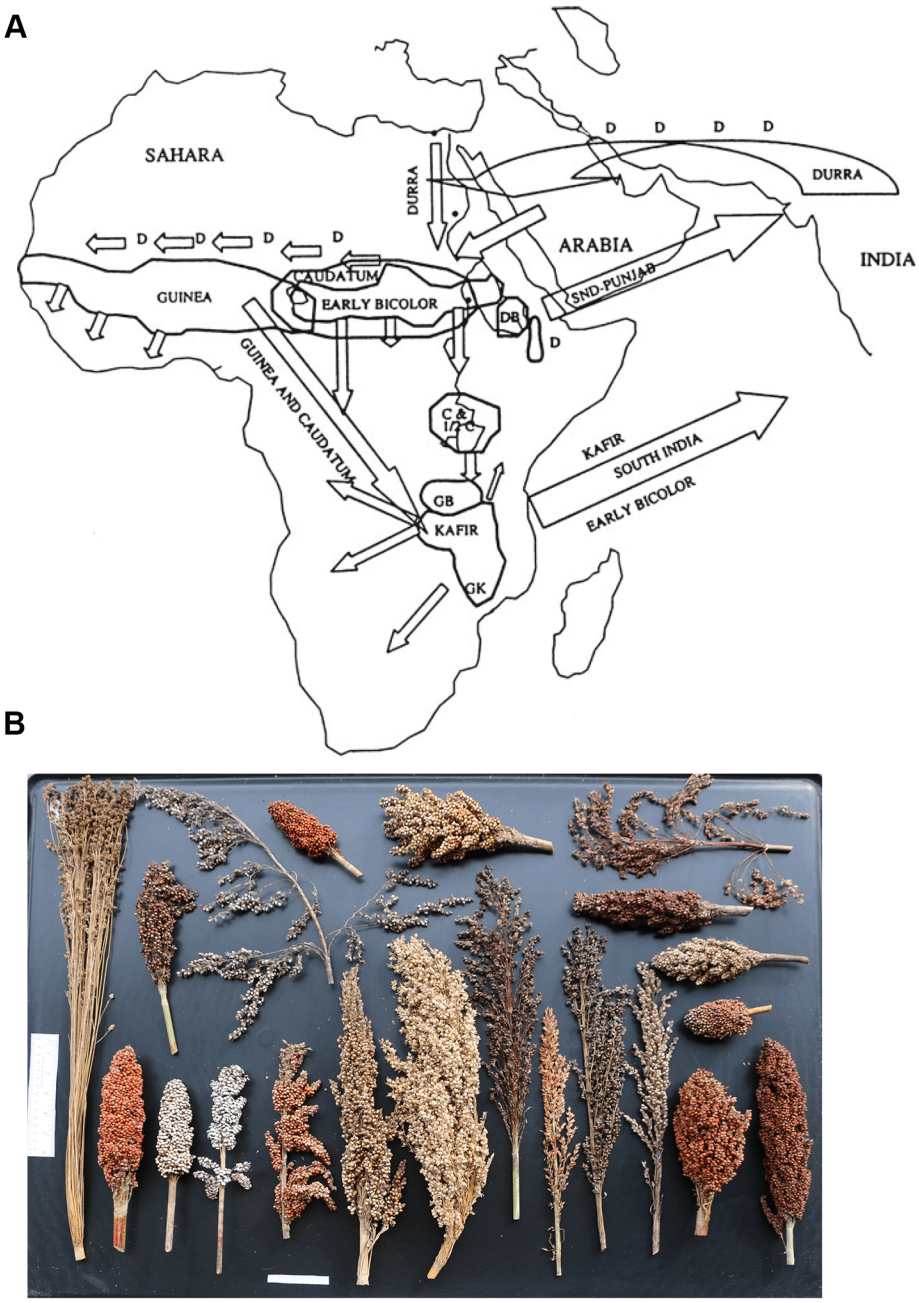


Figure 1.1: A: Possible origin and diversification of sorghum races and possible routes of migration. Adopted from (Kimber et al., 2013). B: Morphological diversity in panicle architecture of sorghum cultivars from different races.

1.2.3 Germplasm, genomic resources and breeding

Sorghum is a diploid species ($2n = 20$) with a sequenced genome consisting of ∼730 Mb of DNA (Paterson et al., 2009). Although some sorghum can have high outcrossing rates (as high as 40%), sorghum is predominantly self-pollinating (Doggett, 1988). The higher outcrossing rates are observed in races, guinea and bicolor, that have open panicles and florets (Djè et al., 2004). The level of linkage disequilibrium (LD) is relatively low compared to other self-pollinated cereal crops, which results in high resolution of LD based genetic mapping almost to the gene level in some genomic regions (Hamblin et al., 2004; Mace et al., 2013; Morris et al., 2013). The community efforts to generate and curate sorghum genomic resources will help us understand the effects of genotype × environment × management, therefore, allowing better utilization of germplasm and genetic resources for crop improvement (Boyles et al., 2019).

The worldwide collection contains over 100,000 sorghum accessions, including approximately 45,000 accessions in the United States Department of Agriculture, National Plant Germplasm System (USDA-NPGS). While these diverse collection of germplasm with extensive genetic potential are still largely untapped, a need to broaden the genetic diversity of US sorghum breeding gene pool was realized during the 1960s in the form of the sorghum conversion program (Stephens et al., 1967). The sorghum conversion program, initiated by the USDA in cooperation with Texas A&M University, has introduced novel genetic variation from exotic tropical germplasm by converting selected tropical genotypes to temperate adapted, photoperiod-insensitive lines with short stature (Adams, 1995). This program created germplasm that are not only early maturing but through introgression of dwarfing loci (*dw*) made sorghum production amenable to mechanical harvesting. The conversion program has been reinstated to convert more tropical sorghum germplasm for US sorghum breeding (Klein et al., 2016). While this ongoing initiative has been the staple source of germplasm for several public and private breeding programs in temperate regions, the conversion of tropical lines through repeated backcrossing is an expensive and labor-intensive process. Therefore, only about 1000-1500 lines have been converted through this program (Bob Klein, personal communication, 2019). Advances in genomic and computational capabilities, however, present new opportunities for effective germplasm screening and selection strategies to exploit novel genetic variation in pre-breeding and population development (Yu et al., 2016).

The cytoplasmic male sterility (CMS) system in sorghum was identified and understood

during the 1950s (Maunder et al., 1959; Stephens et al., 1954). The CMS in sorghum is determined by interactions between the mitochondrial gene in cytoplasm of milo type and nuclear gene of kafir type. This mechanism has been used in breeding and production of hybrid sorghum since the 1960s. The CMS system requires three distinct lines, viz. male sterile female (A), maintainer female (B), and restorer male (R). Since A and B lines have identical nucleus but different cytoplasm, A lines are maintained by crossing using pollen from B lines which lack the fertility restorer genes. Two separate gene pools, female lines (A) and male restorer lines (R), are used for commercial seed production and to exploit the hybrid vigor in sorghum. New A lines are created from genotypes with non-restorer nucleus by introgression of male sterile cytoplasm by crossing with CMS donor and then subsequently backcrossing to the desired genotypes. By early 1960s, over three quarters of the sorghum grown in the US were F_1 hybrids (Maunder, 2000). This technique has made hybrid seed production in sorghum feasible and sorghum probably wouldn't have been a hybrid crop without the CMS. However, effectiveness of CMS is limited by the available R lines and the narrowing of diversity in cytoplasm could have huge repercussion if the male sterile (mostly A1) cytoplasm were to be susceptible to a disease epidemic as happened for Texas cytoplasm in maize (Levings, 1990). One potential alternative to CMS could be the availability and commercial success of effective gametocides for sorghum (Boerman et al., 2019).

Despite these efforts the large amount of genetic diversity is still untapped, and only small proportion of the converted lines are actually used in breeding programs. For example, the ideotype for grain sorghum breeding in the US has predominantly been the intermediate race kafir-caudatum, which is not surprising especially because of the yield potential and panicle architecture of this racial type (Kimber et al., 2013). However, a breeding program today doesn't have to rely on such narrow range of diversity, especially during population development, because of the genomic tools that are available for making selection based on genomic markers.

1.3 Genomics-assisted breeding

Although annual gain in crop productivity has increased over the last century for major cereal crops, the current rate of gain is insufficient considering the rate of population growth and subsequent demand for food in the next couple of decades (Ray et al., 2012). Productivity is affected by various factors including agronomic practices, biotic and abiotic stresses, and cultivar selection.

Genetic gain is the quantitative genetic measure of increase in performance achieved through artificial selection, and is commonly used in animal and plant breeding programs as a measure of annual gain in productivity (Xu et al., 2017). The expected genetic gain is often defined as:

$$\Delta G = \frac{i \times r \times \sigma_A}{t} \tag{1.1}$$

where ΔG is the rate of genetic gain, i is the intensity of selection, r is the accuracy of selection, σ_A is additive genetic variance, and t is time taken per breeding cycle. The genetic gain can be increased by increasing the genetic diversity, increasing the efficiency of selection ($i \times r$), and increasing breeding cycles per year. A recent technique called 'speed breeding' has been shown to increase genetic gain by maximizing number of breeding cycles per year (Li et al., 2018).

The routine use of marker-assisted selection (MAS) in 1990s and early 2000s led to increase in the efficiency of selection in animal and plant breeding, but even the established breeding programs had to use it conscientiously because of high cost of sequencing (Bernardo et al., 2007). Phenotyping, although laborious and time consuming, was the cheaper option for the breeding programs. However, with next generation sequencing the economic burden in genotyping dropped significantly, and as computational capabilities increased single nucleotide polymorphism (SNP) markers became more ubiquitous in genomic analysis (Bernardo, 2008). The ability to generate large number of genome-wide markers led to the rise of new tools like association mapping and genomic selection.

1.3.1 Genome-wide association

Linkage mapping, despite being a powerful method to identify quantitative trait loci (QTL) for a segregating trait, is limited by the amount of allelic diversity and mapping resolution due to limited recombinations within segregating parents (Korte et al., 2013). In contrast, genome-wide association studies (GWAS) have mapped genetic variants associated to phenotypes to a much higher resolution using whole genome markers in a diverse group of individuals (Korte et al., 2013). Application of association studies in plants was realized with development of statistical methods to account for the inherent population structure to control for nonfunctional, spurious associations (Thornsberry et al., 2001). Yu et al. (2006) introduced a unified mixed model approach to account for multiple level of relatedness and showed that it effectively controlled for false positives as well as false negatives. While the population structure is a confounding factor in genotype-phenotype

association, simply controlling for population structure may not always lead to avoidance of false positive or false negative, especially for association analysis of traits that could be strongly correlated to population structure (Lawson et al., 2020; Sul et al., 2018). However, models with increased power to detect genome-wide association without needing to correct for population structure could mitigate those limitations (Klasen et al., 2016; Liu et al., 2016).

While most of the association analysis have been focused on single traits, traits are usually correlated and controlled by genetic loci with pleiotropic effects. Association studies have shown that combined analysis of correlated traits can be effective in detecting additional genetic variants with small effects across multiple traits (Carlson et al., 2019; Korte et al., 2012; Rice et al., 2020; Thoen et al., 2017). Examples of approaches to leverage correlation between traits in association analysis include: the use of ratios of directly related traits in univariate GWAS (Gallagher et al., 2018), combining test statistics from univariate GWAS of each trait to detect pleiotropic effects (Yang et al., 2010), using dimension reduction technique to derive transformed phenotypes for univariate GWAS (Aschard et al., 2014), and directly modeling multiple traits into a multivariate linear mixed models (Korte et al., 2012; Zhou et al., 2014). Furthermore, meta-analyses of results from GWAS have shown promise in linking associated variants to meaningful biological functions in the form of metabolic and biochemical pathways associated with multiple correlated trait (Battenfield et al., 2018; Duarte et al., 2018; Gebreyesus et al., 2019).

1.3.2 Genomic selection/prediction

Most of the traits of interests in plant breeding are complex and controlled by large number of small effect loci. For the traits controlled by several minor effect QTL, association mapping results in poor prediction of line performance because of the biased effect estimates and its inability to detect minor effect QTLs (Jannink et al., 2010). Meuwissen et al. (2001) proposed a technique called genomic selection which mitigates the disadvantages of GWAS by utilizing statistical models that are capable of simultaneous estimation of all marker effects. Genomic selection de-emphasizes the identification of individual polymorphisms for complex traits towards weighing a predicted performance based on model training (Jannink et al., 2010). Genomic selection has potential to increase breeding efficiency by improving genetic gain per selection in a breeding program per unit time. In genomic selection, a 'training population' of individuals, that is both genotyped and phenotyped, is selected to train prediction models that use the genotypic data of untested individuals from the

'candidate population' (prediction set) to produce genomic estimated breeding values (GEBVs). These GEBVs are primarily based on the cumulative effects of markers predicted by the model that was formerly trained, and are used as selection criteria without known function of the underlying genes (Jannink et al., 2010). Cross-validation experiments are often used to calculate the accuracy of prediction models. The prediction accuracy or predictive ability is defined as the correlation between the predicted genetic values and true breeding value, which is often the observed phenotypic values from empirical experiments.

Genomic selection or prediction has created a paradigm shift in plant and animal breeding, and to-date hundreds of studies and applications of this method in crops and livestock for numerous traits have already been reported (Crossa et al., 2017; Meuwissen et al., 2016). In sorghum, genomic prediction studies has been reported for bioenergy traits (Fernandes et al., 2018; Oliveira et al., 2018; Yu et al., 2016), grain yield (Hunt et al., 2018; Sapkota et al., 2020b), drought tolerance (Velazco et al., 2019), and grain composition (Sapkota et al., 2020a). Additionally, Valluru et al. (2019) studied the effect of deleterious variants on genomic prediction of bioenergy traits. Furthermore, Sapkota et al. (2020a) and Velazco et al. (2019) showed increase in accuracy of prediction due to multivariate prediction models over univariate models. The application of genomic prediction can be extended to screening of germplasm and early generation selection during pre-breeding and population improvement (Gaynor et al., 2017; Yu et al., 2016). The large volume of genomic and phenotypic data produced through association studies using diverse individuals across several crop species can be used to train prediction models to evaluate genetic merit of gene bank accessions that have largely remained unused (Yu et al., 2016). Such a strategy is also likely to be advantageous for identifying rare alleles that are more likely to go undetected and purged in pedigreed breeding populations.

The accuracy of genomic prediction models is affected by several factors including training population size, genetic architecture of trait, genetic relatedness, marker density and co-segregation of markers (Combs et al., 2013; Habier et al., 2007, 2013; Zhong et al., 2009). Higher relatedness between individuals in training and test population inflates the overall accuracy of predictions in genomic selection models (Habier et al., 2007). In stratified populations, the population structure can play a vital role in genomic prediction accuracy. In maize and rice diversity panel, Guo et al. (2014b) found that accuracy of prediction is higher when individuals within a subpopulation were used as training population as compared to predictions of a subpopulation using individuals in unrelated

subpopulation for training. Positive relationship between prediction accuracy and heritability or training population size or marker density have been observed across different crops (Combs et al., 2013; Tayeh et al., 2015; Zhang et al., 2017). While model selection is known to have some effect on genomic prediction accuracy, studies have shown that the differences due to model selection are small and insignificant (Crossa et al., 2017; Heslot et al., 2012).

VanRaden (2008) proposed a genomic best linear unbiased prediction (GBLUP) model which has become the most widely used genomic prediction model in animal and plant breeding. This model uses linear predictions as proposed by Henderson (1963) but replaces traditional relationship matrix with genomic relationship matrix that is calculated using bi-allelic marker genotypes and their allele frequencies. Computational research on genomic selection has since emerged with statistical methods capable of incorporating pedigree, genomic, and environmental covariates into statistical-genetic prediction models (Crossa et al., 2017). A major advantage of genomic prediction is the ability to make prediction on new lines and environment based on phenotypic values of available lines in some environments. Extensions of GBLUP model has reported inclusion of genotype $\times$ environment interactions resulting in improvement in prediction accuracy (Burgueño et al., 2012; Jarquín et al., 2014). Genomic prediction models have also been expanded to perform joint analysis of multiple traits using empirical and simulated data (Guo et al., 2014a; Jia et al., 2012). The improvement in prediction accuracy of multi-trait model over single-trait models, however, depends on the heritability of traits and correlation between them (Jia et al., 2012; Lado et al., 2018). Since phenotypic data in breeding programs are collected across multiple environment and for multiple traits, Montesinos-López et al. (2016) developed a Bayesian whole genome prediction model to account for complexity of variance-covariance structure in a combined multi-trait multi-environment (MTME) model. They also developed a computationally efficient Markov Chain Monte Carlo (MCMC) method that produces a full conditional distribution of the parameters leading to an exact Gibbs sampling for the posterior distribution. Increased computational capabilities and affordable sequencing is changing the paradigm in plant and animal breeding. As data-driven research and applications become common place, holistic breeding practices using multi-omics data will become the new normal in this century (Wallace et al., 2018).

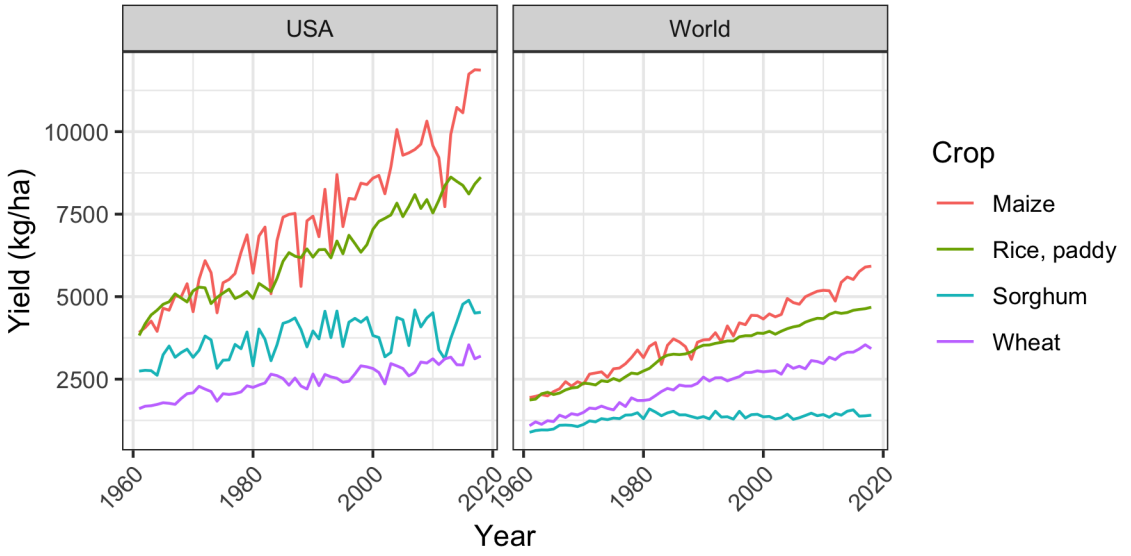


Figure 1.2: Trend of grain yield for major cereal crops. Source: FAOSTAT (2018).

1.4 Grain yield

Yield is the most important trait because it translates directly to economic gain for the farmers. Figure 1.2 shows worldwide and US trends for average grain yield from 1961 to 2018 in major cereal crops. While major gains have been made in wheat, rice and maize, the yields have remained stagnant for sorghum across the world and very little gains has been made even in the US (Figure 1.2). The three-year average yield has not changed for sorghum since the 1980s (FAOSTAT, 2018). Pfeiffer et al. (2019) compared the total genetic gain of 74 hybrids (60 public and 14 commercial) released within a 50-year span and found that yield gain was 8 kg ha^{-1} per year. Yield potential per plant, heterosis, test weight, panicle size, and grain number per panicle were the traits that showed increase, whereas leaf angle, days to maturity, plant height, and yield stability demonstrated little to no change. Overall ~60% of the total yield gains in US sorghum production were attributed to the genetic improvement through sorghum breeding, which is slightly higher than the reported contribution (~50%) of plant breeding towards yield gain in hybrid maize (Duvick, 2005; Pfeiffer et al., 2019). But since the yield gain is so small, the increase in total yield potential was probably marginal compared to that of maize.

Grain yield in cereal crops is a composite trait and is determined by four primary components: planting density, number of panicles per plant, number of grains per panicle, and grain weight

(Boyles et al., 2016; Heinrich et al., 1983). While there might be other morphological and physiological factors that affect grain yield, their effects, however, would be observed through phenotypic changes in grain number or grain weight. Grain number and grain weight are complex traits that are quantitatively inherited and are influenced by both genetic and environmental factors (Austin et al., 1998). These two yield component traits, grain number and grain weight, are not only important for yield potential but also for yield stability (Heinrich et al., 1983). Despite reported evolutionary trade-offs between the grain number and grain weight (reviewed by Sadras, 2007), examples of decoupling the two traits suggests increasing one is possible without decreasing the other (Gambín et al., 2012; Griffiths et al., 2015).

Large proportion of phenotypic variance in yield is attributable to the variance due to environment and genotype × environment, and grain yield is controlled by many genetic loci with small effects (Boyles et al., 2016). This makes grain yield a good target trait for application of genomic selection. Genomic prediction studies in maize (Windhausen et al., 2012), wheat (Saint Pierre et al., 2016), and sorghum (Hunt et al., 2018) have shown potential for increasing genetic gain across multiple environments. Velazco et al. (2019) showed using multi-trait prediction models can increase accuracy of prediction for grain yield by including correlated traits that affect grain yield. These examples suggest the rate of genetic gain for yield can be increased by dissection of yield components, and using grain number and grain weight in trait-assisted and multi-trait genomic prediction (Fernandes et al., 2018).

1.5 Grain composition

Cereal grains are primarily composed of starch, protein, and fat, and combined together these three compositional traits contribute to the total energy provided by the grain (Boyles et al., 2017). Besides, sorghum grains also consists of several health-promoting antioxidants and micronutrients within their grain (Rhodes et al., 2014; Shakoor et al., 2016). Despite success in identification of some genetic variants associated to starch, protein and fat content through genetic mapping, large proportion of genetic effect on phenotypic variance remains unexplained (Boyles et al., 2017; Murray et al., 2008; Rami et al., 1998; Rhodes et al., 2017; Sukumaran et al., 2012). The complex nature of these traits suggest a quantitative inheritance of many small effect loci. Genomic prediction for grain quality traits has previously been reported in crops such as wheat (Battenfield et al., 2018; Haile

et al., 2018), rye (Schulthess et al., 2016), maize (Guo et al., 2014b), and soybean (Duhnen et al., 2017). Studies using near-infrared derived phenotypes in genomic prediction of protein content and end-use quality has shown moderate to high accuracy of prediction in wheat (Battenfield et al., 2018; Hayes et al., 2017). Although grain yield and protein content are known to have negative trade-off, multi-trait genomic prediction models show simultaneous improvement for grain yield and protein content is possible (Haile et al., 2018; Rapp et al., 2018). Since grain composition traits are highly heritable and inter-correlated among each other (Boyles et al., 2017), the use of single-trait as well as multi-trait genomic prediction models can help in rapid genetic gain for these traits.

1.6 Research objectives

The main goal of this study was to implement and assess potential of genomic prediction for grain yield components and grain composition in sorghum. While several empirical as well as simulation studies for genomic prediction are available for cereal crops such as maize and wheat, research and application of genomic prediction in sorghum is lagging. The outcomes of this study provide empirical evidence and methodology for genomics-assisted breeding of sorghum and the genomic regions and candidate genes from significantly associated regions will provide new targets for understanding biology of important grain composition traits: starch and protein.

Objective 1

Examine genetic diversity and population structure, and study their effects on prediction accuracy of grain yield components.

Objective 2

Assess predictability of grain macronutrients and compare multivariate prediction models to univariate models using a recombinant inbred population and a diversity panel.

Objective 3

Conduct univariate and multivariate genome-wide association analysis for starch and protein content, identify potential candidate genes and their interactors, and study the gene expression patterns.

1.7 References

Adams, Sean (1995). "Resetting sorghum's internal clock: program to convert tropical plants is unique among crops". In: *Agricultural Research* 43.12, pp. 18–20.

Aschard, Hugues, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft (2014). "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies". In: *The American Journal of Human Genetics* 94.5, pp. 662–676.

Austin, David F and Michael Lee (1998). "Detection of quantitative trait loci for grain yield and yield components in maize across generations in stress and nonstress environments". In: *Crop Science* 38.5, pp. 1296–1308.

Battenfield, Sarah D, Jaime L Sheridan, Luciano DCE Silva, Kelci J Miclaus, Susanne Dreisigacker, Russell D Wolfinger, Roberto J Peña, Ravi P Singh, Eric W Jackson, Allan K Fritz, et al. (2018). "Breeding-assisted genomics: Applying meta-GWAS for milling and baking quality in CIMMYT wheat breeding program". In: *PloS one* 13.11.

Bernardo, Rex (2008). "Molecular markers and selection for complex traits in plants: learning from the last 20 years". In: *Crop science* 48.5, pp. 1649–1664.

Bernardo, Rex and Jianming Yu (2007). "Prospects for genomewide selection for quantitative traits in maize". In: *Crop Science* 47.3, pp. 1082–1090.

Boerman, Nicholas A, Kyle B Hlavinka, Weixi Zhu, Alan R Dabney, George L Hodnett, and William L Rooney (2019). "Efficacy of the chemical trifluoromethanesulfonamide as a male gametocide in field-grown sorghum [Sorghum bicolor (L.) Moench]". In: *Euphytica* 215.5, p. 96.

Boyles, Richard E, Zachary W Brenton, and Stephen Kresovich (2019). "Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments". In: *The Plant Journal* 97.1, pp. 19–39.

Boyles, Richard E, Elizabeth A Cooper, Matthew T Myers, Zachary Brenton, Bradley L Rauh, Geoffrey P Morris, and Stephen Kresovich (2016). "Genome-wide association studies of grain yield components in diverse sorghum germplasm". In: *The plant genome* 9.2.

Boyles, Richard E, Brian K Pfeiffer, Elizabeth A Cooper, Bradley L Rauh, Kelsey J Zielinski, Matthew T Myers, Zachary Brenton, William L Rooney, and Stephen Kresovich (2017).

"Genetic dissection of sorghum grain quality traits using diverse and segregating populations". In: *Theoretical and applied genetics* 130.4, pp. 697–716.

Brenton, Zachary W, Elizabeth A Cooper, Mathew T Myers, Richard E Boyles, Nadia Shakoor, Kelsey J Zielinski, Bradley L Rauh, William C Bridges, Geoffrey P Morris, and Stephen Kresovich (2016). "A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy". In: *Genetics* 204.1, pp. 21–33.

Burgueño, Juan, Gustavo de los Campos, Kent Weigel, and José Crossa (2012). "Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers". In: *Crop Science* 52.2, pp. 707–719.

Carlson, Maryn O, Gracia Montilla-Bascon, Owen A Hoekenga, Nicholas A Tinker, Jesse Poland, Matheus Baseggio, Mark E Sorrells, Jean-Luc Jannink, Michael A Gore, and Trevor H Yeats (2019). "Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (Avena sativa L.)" In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2963–2975.

Combs, Emily and Rex Bernardo (2013). "Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers". In: *The Plant Genome* 6.1.

Crossa, José, Paulino Pérez-Rodrıguez, Jaime Cuevas, Osval Montesinos-López, Diego Jarquın, Gustavo de los Campos, Juan Burgueño, Juan M González-Camacho, Sergio Pérez-Elizalde, Yoseph Beyene, et al. (2017). "Genomic selection in plant breeding: methods, models, and perspectives". In: *Trends in plant science* 22.11, pp. 961–975.

Dahlberg, Jeff A and Krystyna Wasylikowa (1996). "Image and statistical analyses of early sorghum remains (8000 BP) from the Nabta Playa archaeological site in the Western Desert, southern Egypt". In: *Vegetation History and Archaeobotany* 5.4, pp. 293–299.

Djè, Yao, Myriam Heuertz, Mohamed Ater, Claude Lefèbvre, and Xavier Vekemans (2004). "In situ estimation of outcrossing rate in sorghum landraces using microsatellite markers". In: *Euphytica* 138.3, pp. 205–212.

Doggett, Hugh (1988). *Sorghum*. Longman Group UK Limited.

Duarte, Darlene Ana S, Marina Rufino S Fortes, Marcio de Souza Duarte, Simone EF Guimarães, Lucas L Verardo, Renata Veroneze, André Mauric F Ribeiro, Paulo Sávio Lopes, Marcos Deon V de Resende, and Fabyano Fonseca e Silva (2018). "Genome-wide association studies,

meta-analyses and derived gene network for meat quality and carcass traits in pigs". In: *Animal Production Science* 58.6, pp. 1100–1108.

Duhnen, Alexandra, Amandine Gras, Simon Teyssèdre, Michel Romestant, Bruno Claustres, Jean Daydé, and Brigitte Mangin (2017). "Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy". In: *Crop Science* 57.3, pp. 1325–1337.

Duvick, Donald N (2005). "The contribution of breeding to yield advances in maize (Zea mays L.)" In: *Advances in agronomy* 86, pp. 83–145.

FAOSTAT (2018). "Food and Agriculture Organization of the United Nations, Statistics Division". In: *Economic and Social Development Department, Rome, Italy.* URL: `http://faostat3.fao.org/home/E.%20Accessed:%202020`.

Fernandes, Samuel B, Kaio OG Dias, Daniel F Ferreira, and Patrick J Brown (2018). "Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum". In: *Theoretical and applied genetics* 131.3, pp. 747–755.

Gallagher, Michael D and Alice S Chen-Plotkin (2018). "The post-GWAS era: from association to function". In: *The American Journal of Human Genetics* 102.5, pp. 717–730.

Gambín, Brenda L and Lucas Borrás (2012). "Genotypic diversity in sorghum inbred lines for grain-filling patterns and other related agronomic traits". In: *Crop and Pasture Science* 62.12, pp. 1026–1036.

Gaynor, R Chris, Gregor Gorjanc, Alison R Bentley, Eric S Ober, Phil Howell, Robert Jackson, Ian J Mackay, and John M Hickey (2017). "A two-part strategy for using genomic selection to develop inbred lines". In: *Crop Science* 57.5, pp. 2372–2386.

Gebreyesus, G, AJ Buitenhuis, NA Poulsen, MHPW Visker, Q Zhang, HJF van Valenberg, D Sun, and H Bovenhuis (2019). "Combining multi-population datasets for joint genome-wide association and meta-analyses: The case of bovine milk fat composition traits". In: *Journal of Dairy Science* 102.12, pp. 11124–11141.

Godfray, H Charles J, John R Beddington, Ian R Crute, Lawrence Haddad, David Lawrence, James F Muir, Jules Pretty, Sherman Robinson, Sandy M Thomas, and Camilla Toulmin (2010). "Food security: the challenge of feeding 9 billion people". In: *science* 327.5967, pp. 812–818.

Griffiths, Simon, Luzie Wingen, Julian Pietragalla, Guillermo Garcia, Ahmed Hasan, Daniel Miralles, Daniel F Calderini, Jignaben Bipinchandra Ankleshwaria, Michelle Leverington Waite, James

Simmonds, et al. (2015). “Genetic dissection of grain size and grain number trade-offs in CIMMYT wheat germplasm”. In: *PloS one* 10.3.

Guo, Gang, Fuping Zhao, Yachun Wang, Yuan Zhang, Lixin Du, and Guosheng Su (2014a). “Comparison of single-trait and multiple-trait genomic prediction models”. In: *BMC genetics* 15.1, p. 30.

Guo, Zhigang, Dominic M Tucker, Christopher J Basten, Harish Gandhi, Elhan Ersoz, Baohong Guo, Zhanyou Xu, Daolong Wang, and Gilles Gay (2014b). “The impact of population structure on genomic prediction in stratified populations”. In: *Theoretical and applied genetics* 127.3, pp. 749–762.

Habier, David, Rohan L Fernando, and Jack CM Dekkers (2007). “The impact of genetic relationship information on genome-assisted breeding values”. In: *Genetics* 177.4, pp. 2389–2397.

Habier, David, Rohan L Fernando, and Dorian J Garrick (2013). “Genomic BLUP decoded: a look into the black box of genomic prediction”. In: *Genetics* 194.3, pp. 597–607.

Haile, Jemanesh K, Amidou N’Diaye, Fran Clarke, John Clarke, Ron Knox, Jessica Rutkoski, Filippo M Bassi, and Curtis J Pozniak (2018). “Genomic selection for grain yield and quality traits in durum wheat”. In: *Molecular breeding* 38.6, p. 75.

Hamblin, Martha T, Sharon E Mitchell, Gemma M White, Javier Gallego, Rakesh Kukatla, Rod A Wing, Andrew H Paterson, and Stephen Kresovich (2004). “Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of Sorghum bicolor”. In: *Genetics* 167.1, pp. 471–483.

Harlan, Jack R and Ann Stemler (1976). *The races of sorghum in Africa*. Mouton: The Hague, The Netherlands Paris, France, pp. 465–478.

Harlan, JR and JMJ De Wet (1972). “A simplified classification of cultivated sorghum 1”. In: *Crop science* 12.2, pp. 172–176.

Hayes, BJ, J Panozzo, CK Walker, AL Choy, S Kant, D Wong, J Tibbits, HD Daetwyler, S Rochfort, MJ Hayden, et al. (2017). “Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes”. In: *Theoretical and applied genetics* 130.12, pp. 2505–2519.

Heinrich, GM, CA Francis, and JD Eastin (1983). “Stability of Grain Sorghum Yield Components Across Diverse Environments 1”. In: *Crop Science* 23.2, pp. 209–212.

Henderson, Charles R (1963). "Selection index and expected genetic advance". In: *Statisitical Genetics and Plant Breeding*.

Heslot, Nicolas, Hsiao-Pei Yang, Mark E Sorrells, and Jean-Luc Jannink (2012). "Genomic selection in plant breeding: a comparison of models". In: *Crop science* 52.1, pp. 146–160.

Hunt, Colleen H, Fred A van Eeuwijk, Emma S Mace, Ben J Hayes, and David R Jordan (2018). "Development of genomic prediction in sorghum". In: *Crop Science* 58.2, pp. 690–700.

Jannink, Jean-Luc, Aaron J Lorenz, and Hiroyoshi Iwata (2010). "Genomic selection in plant breeding: from theory to practice". In: *Briefings in functional genomics* 9.2, pp. 166–177.

Jarquín, Diego, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. (2014). "A reaction norm model for genomic selection using high-dimensional genomic and environmental data". In: *Theoretical and applied genetics* 127.3, pp. 595–607.

Jia, Yi and Jean-Luc Jannink (2012). "Multiple-trait genomic selection methods increase genetic value prediction accuracy". In: *Genetics* 192.4, pp. 1513–1522.

Khush, Gurdev S (2001). "Green revolution: the way forward". In: *Nature reviews genetics* 2.10, pp. 815–822.

Kimber, Clarissa T, Jeff A Dahlberg, and Stephen Kresovich (2013). "The gene pool of Sorghum bicolor and its improvement". In: *Genomics of the Saccharinae*. Springer, pp. 23–41.

Klasen, Jonas R, Elke Barbez, Lukas Meier, Nicolai Meinshausen, Peter Bühlmann, Maarten Koornneef, Wolfgang Busch, and Korbinian Schneeberger (2016). "A multi-marker association method for genome-wide association studies without the need for population structure correction". In: *Nature communications* 7.1, pp. 1–8.

Klein, RR, FR Miller, S Bean, and PE Klein (2016). "Registration of 40 converted germplasm sources from the reinstated sorghum conversion program". In: *Journal of Plant Registrations* 10.1, pp. 57–61.

Korte, Arthur and Ashley Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1, p. 29.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). "A mixed-model approach for genome-wide association studies of correlated traits in structured populations". In: *Nature genetics* 44.9, p. 1066.

Lado, Bettina, Daniel Vázquez, Martin Quincke, Paula Silva, Ignacio Aguilar, and Lucia Gutiérrez (2018). "Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality". In: *Theoretical and Applied Genetics* 131.12, pp. 2719–2731.

Lawson, Daniel John, Neil Martin Davies, Simon Haworth, Bilal Ashraf, Laurence Howe, Andrew Crawford, Gibran Hemani, George Davey Smith, and Nicholas John Timpson (2020). "Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?" In: *Human genetics* 139.1, pp. 23–41.

Levings, Charles S (1990). "The Texas cytoplasm of maize: cytoplasmic male sterility and disease susceptibility". In: *Science* 250.4983, pp. 942–947.

Li, Huihui, Awais Rasheed, Lee T Hickey, and Zhonghu He (2018). "Fast-forwarding genetic gain". In: *Trends in plant science* 23.3, pp. 184–186.

Liu, Xiaolei, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang (2016). "Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies". In: *PLoS genetics* 12.2.

Mace, Emma S, Shuaishuai Tai, Edward K Gilding, Yanhong Li, Peter J Prentis, Lianle Bian, Bradley C Campbell, Wushu Hu, David J Innes, Xuelian Han, et al. (2013). "Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum". In: *Nature communications* 4, p. 2320.

Maunder, AB (2000). *History of cultivar development in the United States: From memoirs of AB Maunder-sorghum breeder*, pp. 191–223.

Maunder, AB and RC Pickett (1959). "The Genetic Inheritance of Cytoplasmic-Genetic Male Sterility in Grain Sorghum 1". In: *Agronomy journal* 51.1, pp. 47–49.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps". In: *Genetics* 157.4, pp. 1819–1829. ISSN: 0016-6731. URL: `https://www.genetics.org/content/157/4/1819`.

Meuwissen, Theo, Ben Hayes, and Mike Goddard (2016). "Genomic selection: A paradigm shift in animal breeding". In: *Animal frontiers* 6.1, pp. 6–14.

Montesinos-López, Osval A, Abelardo Montesinos-López, José Crossa, Fernando H Toledo, Oscar Pérez-Hernández, Kent M Eskridge, and Jessica Rutkoski (2016). "A genomic Bayesian

multi-trait and multi-environment model". In: *G3: Genes, Genomes, Genetics* 6.9, pp. 2725–2744.

Morris, Geoffrey P, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. (2013). "Population genomic and genome-wide association studies of agroclimatic traits in sorghum". In: *Proceedings of the National Academy of Sciences* 110.2, pp. 453–458.

Murray, Seth C, Arun Sharma, William L Rooney, Patricia E Klein, John E Mullet, Sharon E Mitchell, and Stephen Kresovich (2008). "Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates". In: *Crop Science* 48.6, pp. 2165–2179.

Oliveira, Amanda Avelar de, Maria Marta Pastina, Rafael Augusto da Costa Parrella, Roberto Willians Noda, Maria Lúcia Ferreira Simeone, Robert Eugene Schaffert, Jurandir Vieira de Magalhães, Cynthia Maria Borges Damasceno, Gabriel Rodrigues Alves Margarido, et al. (2018). "Genomic prediction applied to high-biomass sorghum for bioenergy production". In: *Molecular Breeding* 38.4, p. 49.

Paterson, Andrew H, John E Bowers, Remy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Uffe Hellsten, Therese Mitros, Alexander Poliakov, et al. (2009). "The Sorghum bicolor genome and the diversification of grasses". In: *Nature* 457.7229, pp. 551–556.

Pfeiffer, Brian K, Dennis Pietsch, Ronnie W Schnell, and William L Rooney (2019). "Long-Term Selection in Hybrid Sorghum Breeding Programs". In: *Crop Science* 59.1, pp. 150–164.

Rami, J-F, Philippe Dufour, Gilles Trouche, Geneviève Fliedel, Christian Mestres, Fabrice Davrieux, P Blanchard, and Perla Hamon (1998). "Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (Sorghum bicolor L. Moench)". In: *Theoretical and applied genetics* 97.4, pp. 605–616.

Rapp, M, V Lein, F Lacoudre, J Lafferty, E Müller, G Vida, V Bozhanova, A Ibraliu, P Thorwarth, HP Piepho, et al. (2018). "Simultaneous improvement of grain yield and protein content in durum wheat by different phenotypic indices and genomic selection". In: *Theoretical and applied genetics* 131.6, pp. 1315–1329.

Ray, Deepak K, Navin Ramankutty, Nathaniel D Mueller, Paul C West, and Jonathan A Foley (2012). "Recent patterns of crop yield growth and stagnation". In: *Nature communications* 3.1, pp. 1–7.

Rhodes, Davina H, Leo Hoffmann Jr, William L Rooney, Punna Ramu, Geoffrey P Morris, and Stephen Kresovich (2014). "Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm". In: *Journal of agricultural and food chemistry* 62.45, pp. 10916–10927.

Rhodes, Davina H, Leo Hoffmann, William L Rooney, Thomas J Herald, Scott Bean, Richard Boyles, Zachary W Brenton, and Stephen Kresovich (2017). "Genetic architecture of kernel composition in global sorghum germplasm". In: *BMC genomics* 18.1, p. 15.

Rice, Brian R, Samuel B Fernandes, and Alexander E Lipka (2020). "Multi-Trait Genome-wide Association Studies Reveal Loci Associated with Maize Inflorescence and Leaf Architecture". In: *Plant and Cell Physiology*.

Sadras, Victor O (2007). "Evolutionary aspects of the trade-off between seed size and number in crops". In: *Field Crops Research* 100.2-3, pp. 125–138.

Saint Pierre, C, Juan Burgueño, Jose Crossa, G Fuentes Dávila, P Figueroa López, E Solıs Moya, J Ireta Moreno, VM Hernández Muela, VM Zamora Villa, P Vikram, et al. (2016). "Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones". In: *Scientific reports* 6, p. 27312.

Sapkota, Sirjan, J Lucas Boatwright, Kathleen Elizabeth Jordan, Richard Boyles, and Stephen Kresovich (2020a). "Multi-trait regressor stacking increased genomic prediction accuracy of sorghum grain composition". In: *bioRxiv*.

Sapkota, Sirjan, Richard Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich (2020b). "Impact of sorghum racial structure and diversity on genomic prediction of grain yield components". In: *Crop Science*.

Schulthess, Albert Wilhelm, Yu Wang, Thomas Miedaner, Peer Wilde, Jochen C Reif, and Yusheng Zhao (2016). "Multiple-trait and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes". In: *Theoretical and Applied Genetics* 129.2, pp. 273–287.

Shakoor, Nadia, Greg Ziegler, Brian P Dilkes, Zachary Brenton, Richard Boyles, Erin L Connolly, Stephen Kresovich, and Ivan Baxter (2016). "Integration of experiments across diverse en-

vironments identifies the genetic determinants of variation in Sorghum bicolor seed element composition". In: *Plant physiology* 170.4, pp. 1989–1998.

Stephens, JC and RF Holland (1954). "Cytoplasmic male sterility for hybrid sorghum seed production". In: *Agron. J* 46.1, pp. 20–23.

Stephens, JC, FR Miller, and DT Rosenow (1967). "Conversion of alien Sorghums to early combine genotypes 1". In: *Crop Science* 7.4, pp. 396–396.

Sukumaran, Sivakumar, Wenwen Xiang, Scott R Bean, Jeffrey F Pedersen, Stephen Kresovich, Mitchell R Tuinstra, Tesfaye T Tesso, Martha T Hamblin, and Jianming Yu (2012). "Association mapping for grain quality in a diverse sorghum collection". In: *The Plant Genome* 5.3, pp. 126–135.

Sul, Jae Hoon, Lana S Martin, and Eleazar Eskin (2018). "Population structure in genetic studies: Confounding factors and mixed models". In: *PLoS genetics* 14.12.

Swigoňová, Zuzana, Jinsheng Lai, Jianxin Ma, Wusirika Ramakrishna, Victor Llaca, Jeffrey L Bennetzen, and Joachim Messing (2004). "Close split of sorghum and maize genome progenitors". In: *Genome research* 14.10a, pp. 1916–1923.

Tayeh, Nadim, Anthony Klein, Marie-Christine Le Paslier, Françoise Jacquin, Hervé Houtin, Céline Rond, Marianne Chabert-Martinello, Jean-Bernard Magnin-Robert, Pascal Marget, Grégoire Aubert, et al. (2015). "Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy". In: *Frontiers in plant science* 6, p. 941.

Taylor, John RN, Tilman J Schober, and Scott R Bean (2006). "Novel food and non-food uses for sorghum and millets". In: *Journal of cereal science* 44.3, pp. 252–271.

Thoen, Manus PM, Nelson H Davila Olivas, Karen J Kloth, Silvia Coolen, Ping-Ping Huang, Mark GM Aarts, Johanna A Bac-Molenaar, Jaap Bakker, Harro J Bouwmeester, Colette Broekgaarden, et al. (2017). "Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping". In: *New Phytologist* 213.3, pp. 1346–1362.

Thornsberry, Jeffry M, Major M Goodman, John Doebley, Stephen Kresovich, Dahlia Nielsen, and Edward S Buckler (2001). "Dwarf8 polymorphisms associate with variation in flowering time". In: *Nature genetics* 28.3, pp. 286–289.

Valluru, Ravi, Elodie E Gazave, Samuel B Fernandes, John N Ferguson, Roberto Lozano, Pradeep Hirannaiah, Tao Zuo, Patrick J Brown, Andrew DB Leakey, Michael A Gore, et al. (2019).

"Deleterious mutation burden and its association with complex traits in sorghum (Sorghum bicolor)". In: *Genetics* 211.3, pp. 1075–1087.

VanRaden, Paul M (2008). "Efficient methods to compute genomic predictions". In: *Journal of dairy science* 91.11, pp. 4414–4423.

Velazco, Julio G, David R Jordan, Emma S Mace, Colleen H Hunt, Marcos Malosetti, and Fred A Van Eeuwijk (2019). "Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis". In: *Frontiers in plant science* 10.

Wallace, Jason G, Eli Rodgers-Melnick, and Edward S Buckler (2018). "On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics". In: *Annual review of genetics*.

Wendorf, Fred, Angela E Close, Romuald Schild, Krystyna Wasylikowa, Rupert A Housley, Jack R Harlan, and Halina Królik (1992). "Saharan exploitation of plants 8,000 years BP". In: *Nature* 359.6397, pp. 721–724.

Windhausen, Vanessa S, Gary N Atlin, John M Hickey, Jose Crossa, Jean-Luc Jannink, Mark E Sorrells, Babu Raman, Jill E Cairns, Amsal Tarekegne, Kassa Semagn, et al. (2012). "Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments". In: *G3: Genes, Genomes, Genetics* 2.11, pp. 1427–1436.

Wolfe, Kenneth H, Manolo Gouy, Yau-When Yang, Paul M Sharp, and When-Hsiung Li (1989). "Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data". In: *Proceedings of the National Academy of Sciences* 86.16, pp. 6201–6205.

Xu, Yunbi, Ping Li, Cheng Zou, Yanli Lu, Chuanxiao Xie, Xuecai Zhang, Boddupalli M Prasanna, and Michael S Olsen (2017). "Enhancing genetic gain in the era of molecular breeding". In: *Journal of Experimental Botany* 68.11, pp. 2641–2666.

Yang, Qiong, Hongsheng Wu, Chao-Yu Guo, and Caroline S Fox (2010). "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests". In: *Genetic epidemiology* 34.5, pp. 444–454.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". In: *Nature genetics* 38.2, pp. 203–208.

Yu, Xiaoqing, Xianran Li, Tingting Guo, Chengsong Zhu, Yuye Wu, Sharon E Mitchell, Kraig L Roozeboom, Donghai Wang, Ming Li Wang, Gary A Pederson, et al. (2016). "Genomic prediction contributing to a promising global strategy to turbocharge gene banks". In: *Nature Plants* 2.10, pp. 1–7.

Zhang, Ao, Hongwu Wang, Yoseph Beyene, Kassa Semagn, Yubo Liu, Shiliang Cao, Zhenhai Cui, Yanye Ruan, Juan Burgueño, Felix San Vicente, et al. (2017). "Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations". In: *Frontiers in plant science* 8, p. 1916.

Zhong, Shengqiang, Jack CM Dekkers, Rohan L Fernando, and Jean-Luc Jannink (2009). "Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study". In: *Genetics* 182.1, pp. 355–364.

Zhou, Xiang and Matthew Stephens (2014). "Efficient multivariate linear mixed model algorithms for genome-wide association studies". In: *Nature methods* 11.4, p. 407.

Zhu, Fan (2014). "Structure, physicochemical properties, modifications, and uses of sorghum starch". In: *Comprehensive Reviews in Food Science and Food Safety* 13.4, pp. 597–610.

Chapter 2

Impact of Sorghum Racial Structure and Diversity on Genomic Prediction of Grain Yield Components

Citation: Sapkota, S., Boyles, R., Cooper, E., Brenton, Z., Myers, M., and Kresovich, S. Impact of sorghum racial structure and diversity on genomic prediction of grain yield components. *Crop Science*. 2020; 60: 132– 148. https://doi.org/10.1002/csc2.20060.

Supplementary file: Appendix A

Impact of sorghum racial structure and diversity on genomic prediction of grain yield components

Sirjan Sapkota,* Richard Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich

Affiliations: S. Sapkota, Z. Brenton, M. Myers and S. Kresovich, Advanced Plant Technology Program, Clemson University, Clemson, SC 29634; S. Sapkota, R. Boyles, and S. Kresovich, Department of Plant and Environmental Sciences, Clemson University, Clemson SC 29634; R. Boyles, Pee Dee Research and Education Center, Clemson University, Florence, SC 29506; E. Cooper, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223.

*Corresponding author (ssapkot@g.clemson.edu).

Abbreviations: AR, across race; BL, terminal branch length; BLUE, best linear unbiased estimator; CV, cross validation; DTA, days to anthesis; FLH, flag leaf height; GEBV, genomic estimated breeding values; GN, grain number; GW, grain weight; GY, grain yield; LD, linkage disequilibrium; MAF, minor allele frequency; PCA, principal component analysis; PH, plant height; PL, panicle length; GBLUP: genomic best linear unbiased prediction; *r*: prediction accuracy; SRT, single race training; WR, within race.

Abstract

Population structure is an important factor that affects the accuracy of estimated breeding values in genomic prediction. Natural sorghum populations exhibit population structure resulting from genetic and morphological differentiation due to evolutionary divergence. To study the impact of sorghum racial structure and diversity in genomic prediction, we conducted two cross validation (CV) experiments: CV1; proportional sampling from races, and 2) CV2; sampling from across race (AR) or within race (WR). A diversity panel with 389 individuals with 224,007 single nucleotide polymorphisms were used for genomic prediction. Genomic heritabilities for traits were positively correlated (0.63) with their mean prediction accuracy (*r)* from CV1, and within-subpopulation variance accounted for about 80% of total genetic variance. CV1 prediction accuracy ranged from 0.52 to 0.69, but *r* declined by 39% and 54% on average for WR and AR methods, respectively. As a predictor race explained 30 to 50% of covariance for grain and panicle traits but race was a bad predictor of plant height, as expected. Grain weight was consistently the best predicted trait across CV1 and CV2 methods except in AR. Difference in average *r* for WR and AR was greater in durra and caudatum, small in kafir, and non-existent in guinea and mixed. We observed higher prevalence of minor alleles among guinea and mixed subgroups highlighting contribution of allelic diversity towards prediction accuracy. Genomic prediction in sorghum will benefit from utilization of inter-racial diversity and we emphasize the need for further investigations into the role of racial structure in genomic prediction.

Introduction

Cultivated crops have undergone genetic bottlenecks as a result of domestication and artificial selection. Genetic diversity in modern crops is further reduced by current plant breeding practices because most of the cultivars are derived from genetically-related varieties that represent a very small fraction of the global diversity of plant germplasm for any species (McCouch, et al., 2013). Effective utilization of genetic diversity to increase resilience and crop yield potential will remain a key aspect to meet the projected food demands in the next few decades and reduce vulnerability to biotic and abiotic stresses,.

Sorghum is an important cereal crop grown and consumed as a staple by over half a billion people in the semi-arid tropics. The earliest known record of sorghum seeds are the charred remains from 8000 years before present found at Nabta Playa near the Egyptian-Sudanese border during archeological excavations (Wendorf, et al., 1992). After early domestication likely near the Sahel region of sub-Saharan Africa, further migration and adaptation of early sorghum domesticates occurred across Africa and Asia. Those demographic events led to the evolution of morphologically and geographically diverse groups that are classified into five major races and 10 intermediate races of sorghum (Harlan and de Wet, 1972, Harlan and Stemler, 1976). This phenotype-based classification of sorghum races has been supported by genetic evidences in a global diversity panel (Brown, et al., 2011). Furthermore, the linkage disequilibrium (LD) in sorghum has shown presence of strong genetic bottleneck and patterns of disruptive selection across the sorghum genome as a result of domestication, adaptation, and diversifying selection (Mace, et al., 2013, Morris, et al., 2013, Wang, et al., 2013).

The sorghum conversion program (SCP) was initiated by the United States Department of Agriculture (USDA) in cooperation with Texas A&M University to introduce novel genetic variation from exotic tropical germplasm by converting selected tropical genotypes to temperate adapted, photoperiod-insensitive lines with short stature (Stephens et al. 1967). This ongoing initiative has been the staple source of germplasm for several public and private breeding programs in temperate regions. However, the conversion of tropical lines through repeated backcrossing is an expensive and labor-intensive process. Therefore, only 1000-1500 tropical lines have been converted, and these converted lines represent a limited

fraction of USDA and worldwide collection of sorghum germplasm (Bob Klein, personal communication). Recent advances in genomic and computational capabilities present opportunities for identifying effective strategies for introducing and screening of germplasm for novel genetic variation, which can benefit breeders by making selection of prebreeding germplasm more accurate and meaningful.

Genomic prediction (also known as genomic selection or genome-wide selection) is a method to simultaneously estimate the effects of all genetic markers and use those marker effects to estimate breeding values (Meuwissen et al. 2001; Bernardo and Yu 2007). The marker effects are estimated using both genotypic and phenotypic data from a training population, which can then be used to predict genomic estimated breeding value (GEBV) using only the genotypic information in a testing population. The accuracy of prediction is measured as the correlation between GEBVs and true genetic values, often represented by observed phenotypic values. Genomic prediction is usually applied in breeding populations where the training and testing population have a shared pedigree, but its application can be extended to screening and selection for pre-breeding or population improvement (Yu et al. 2016; Gaynor et al. 2017). Every year, an increasingly larger number of association studies are conducted for allele mining by breeding and genetics programs across the world. Large-volume phenotypic datasets generated from these studies can be applied in the investigation and application of genomic prediction models across ranges of crops and traits. These resources can then be utilized in careful strategies to tap into the large number of gene bank accessions by screening for useful genetic variation with potential to enhance genetic gain (Yu et al. 2016).

The implementation of genomic prediction in a diverse and stratified population, however, requires careful consideration of the genomic relationship and population structure (Jannink et al. 2010; Habier et al. 2007). Population structure has been shown to affect the accuracy of genomic prediction in stratified populations across several crop species (Guo, et al., 2014, Ly, et al., 2013, Norman, et al., 2018). Population structure analysis can be done using non-model based approaches like principal component analysis (Patterson et al. 2006; Price et al. 2006) or model based clustering approaches like ADMIXTURE (Alexander et al. 2009). Incorporating population structure estimates into both association and prediction studies has proven useful, but the methods for including population structure in the model can vary. While the inclusion of population structure as a covariate has been successfully applied in mixed models for association studies,

the use of population structure as fixed effects in genomic best linear unbiased prediction (GBLUP) models would be concerning because they already enter into the model as random effects (de los Campos and Sorensen, 2014, Janss, et al., 2012, Price, et al., 2010). One approach to account for population structure in genomic prediction is designing a cross validation scheme that ensures equal representation of each subpopulation in training and validation sets (Albrecht, et al., 2011, Guo, et al., 2014). Alternatively, in order to avoid biased estimation due to the presence of genetic structure and familial relatedness, prediction analysis can be performed by partitioning the genomic variability into within and across group components (Guo, et al., 2014, Norman, et al., 2018, Technow, et al., 2012). Because of distinct racial structure in sorghum, an approach to account for contribution of racial structure in prediction accuracy by decomposing variance-covariance components into expectations due to race and covariance from individuals within a race could be beneficial.

A recent simulation study has highlighted the advantages of genomics-assisted recurrent selection over phenotypic recurrent selection in a nascent and small sorghum breeding program, emphasizing the need for further investigations on genomic selection in sorghum (Muleta, et al., 2019). Since inter-racial diversity is important for heterotic gain, application of genomic prediction in sorghum breeding will benefit from investigations into the effect of racial structure on prediction accuracy. While the effect of population structure in genomic prediction has been extensively studied in major cereal crops, the distinctive evolutionary history and racial structure of sorghum merits the need for investigation into the effect of sorghum racial structure in genomic prediction (Guo, et al., 2014, Isidro, et al., 2015, Norman, et al., 2018). A previous study examined the effect of genetic relatedness on genomic prediction of pedigreed male lines in a sorghum breeding program, however, there has been no studies on the role of sorghum racial structure in genomic prediction (Hunt, et al., 2018). The objectives of our study were to estimate genetic structure and diversity among sorghum races and implement genomic prediction for plant architecture and grain yield traits to examine the effect of racial structure on prediction accuracy using a grain diversity panel.

Materials and methods

Plant materials, field design, and phenotyping

The sorghum diversity panel used in this study consisted of 389 diverse sorghum accessions, including 332 accessions from the United States sorghum association panel (SAP) developed by Casa, et al. (2008). Additional accessions were included for diversity and elite grain characteristics (Boyles, et al., 2016). The population was planted in randomized complete block design with two replications in 2013, 2014, and 2017 field seasons at the Clemson University Pee Dee Research and Education Center in Florence, South Carolina. The accessions were assigned to blocks within the replication based on height, maturity, and photoperiod sensitivity. Each plot was two 6.1 m rows spaced 0.726 m apart with a targeted planting density of 130,000 plants ha^{-1} assuming 75% plant establishment rate. In 2017, an average plant density of ~62,350 plants ha^{-1} was calculated based on plant stand count and row length at 24 days after planting (DAP). Fields were irrigated when plants showed signs of stress in order to avoid confounding effects of maturity and varying degree of drought tolerance in the population on yield. The details on agronomic practices for 2013 and 2014 can be found in Boyles, et al. (2016). In 2017, a lay-by of 93 Kg ha^{-1} N was applied at 35 DAP in addition to variable rate of fertilizer (N, P, K) applications before planting. Preemergence and postemergence herbicide applications in 2017 were consistent with 2013 and 2014. A single application of 0.5 L ha^{-1} of Sivanto^{TM} Prime (Bayer CropScience) was administered at 60 DAP to control sugarcane aphid population.

Three consecutive plants from the odd row of each plot was selected for phenotyping in order to prevent biases due to row effect. We also avoided plants from beginning and end of the row to account for border effect. The detail procedures for phenotyping of agronomic and grain phenotypes has previously been described in Boyles et. al. (2016; 2017) . Days to anthesis (DTA) for each plot was measured as the number of days from planting to when 50% of the plants in the plot were at mid-bloom. Plant height (PH) was measured from ground to the apex of the primary panicle at physiological maturity. Flag leaf height (FLH), panicle length (PL), and terminal branch length (BL) of each plant harvested in 2017 were measured. Flag leaf height was measured as the height from ground to the flag leaf of the plant. Panicle length was measured as length of primary panicle from the terminal branch to the apex of the panicle, and BL was measured as

length of the two terminal primary branches, respectively. A more detailed description of these inflorescence architecture traits can be found in Brown, et al. (2006). Grain yield components were phenotyped from primary panicle of three consecutive plants harvested at physiological maturity as previously mentioned. Panicles were air dried to a constant moisture (10-12%) before threshing. Threshed seed were processed through seed counters (Old Mill Model 900-2) to measure grain number per primary panicle (GN). Grain yield per primary panicle (GY) in grams (g) was measured with a Discovery series scale (Ohaus). Subsequently, thousand grain weight (GW) was calculated from GY and GN; GW (g) = (GY/GN) × 1000.

Phenotypic analysis

R statistical software was used for phenotypic analysis (R Development Core Team, 2016). Simple mean of phenotypic values were calculate for each replication with the years. The phenotypic means of the traits were fitted into a linear mixed model analysis using lme4 package in R (Bates, et al., 2015). We fit the following mixed model equation:

$$y_{ijk} = \mu + G_i + E_j + G_i E_j + R_k(E_j) + e_{ijk} \quad (1)$$

where y_{ijk} is the phenotypic value for genotype i, year j, and replication k within the year j; μ is the population mean; G_i is the fixed effects of i^{th} genotype; E_j, GE_{ij}, $R_k(E_j)$ are random effects of year, genotype × year, and replication within the year, respectively; and e_{ijk} is the random effect of residuals, with $e \sim N(0, \sigma_e^2)$. Since phenotypes for PL, BL and FLH were only available from the year 2017, the model was fit with just the random effect of replications within the year and fixed effect of genotypes. Best linear unbiased estimates (BLUEs) for the traits were calculated from the fixed effect of the genotypes. Correlation plots and histograms for the estimated phenotypic means were generated using the *pairs.panels* function within the R package Psych (Revelle, 2011).

Genotyping

Genetic characterization of the diversity panel was done using genotyping-by-sequencing (GBS) as previously described in Morris, et al. (2013). Sequenced reads were aligned to the sorghum reference assembly (BTx623 v3.1, www.phytozome.net) using burrow-wheelers aligner (Li and Durbin 2010). SNP calling, imputation and filtering were done using the TASSEL 5.0 pipeline (Glaubitz, et al., 2014). A total of 515,318 SNPs was called and subsequently imputed using the *FILLINFindHaplotypesPlugin* and *FILLINImputationPlugin* in TASSEL. FILLIN (Fast, Inbred Line Library ImputatioN) imputes missing genotypes by: (1) haplotype generation using inbred segments that share identity by state, and (2) imputation of resulting haplotypes back onto the target samples (Swarts, et al., 2014). SNP sites were filtered to remove sites with a minor allele frequency (MAF) < 0.01, and sites present in at least 90% of the individuals were retained. A final SNP matrix with a total of 224,007 SNPs was created and used for subsequent genomic analysis and predictions. In the final SNP matrix, all genotypes had less than 10% missing sites. The final SNP genotype matrix was further filtered to retain SNPs with MAF > 0.05 for estimation of the decay of linkage disequilibrium (LD).

Population structure and genetic diversity

The final SNP matrix was used to first identify the optimum number of clusters based on cross validation of error, then used to calculate the ancestry coefficients using ADMIXTURE (Alexander, et al., 2009). Admixture ancestry coefficients (Q matrix) were estimated using the default block relaxation algorithm. We also calculated covariances for the first five principal components (PCs) using the SNP data in TASSEL. A common subset of 35,277 SNPs between the diversity panel and *S. propinquum* from Mace, et al. (2013) was used for neighbor joining tree estimation using the *Cladogram* function in TASSEL. This function first calculates distance between each pair of taxa using modified Euclidean distance (homozygote is 100% similar to itself and heterozygote is 50% similar to itself) and then estimates tree using neighbor joining algorithm (Glaubitz, et al., 2014). The tree was visualized using the web based software Interactive Tree of Life (Letunic and Bork, 2016).

Nucleotide diversity (θ_π), Tajima's D, and genetic differentiation (F_{st}) was estimated from the final SNP matrix with the *vcftools* program (Danecek, et al., 2011) using a non-overlapping sliding window of 100 kb. The window size of 100 kb was chosen to avoid sampling error that could arise from variabilities in SNP marker distribution when low coverage sequencing is used for genotyping (Gusnanto, et al., 2014). Minor allele frequency for each SNP site was also calculated using *vcftools*. Linkage disequilibrium was calculated for pairs of alleles using a sliding window of 50 SNPs in TASSEL, and decay of LD with distance was evaluated using non-linear regression *nls* function in R (R Development Core Team, 2016) with a maximum iteration of 100. Expected values of squared allele-frequency correlation (r^2) under drift equilibrium was calculated using the equation from Hill and Weir (1988) as explained in Remington, et al. (2001). Then, average r^2 and average LD half decay distance (bp) were calculated. Nei's expected heterozygosity was calculated on a per locus basis using the *heterozygosity* function in the R package Pegas (Nei, 1987, Paradis, 2010). Average expected heterozygosity was calculated as mean of heterozygosity across all polymorphic sites.

Genomic prediction and heritability

Statistical model for prediction

Genomic best linear unbiased prediction (GBLUP) model was implemented using *kin.blup* function in R package rrBLUP (Endelman, 2011). In the GBLUP model:

$$y = \mu + g + e \quad (2)$$

y is a vector of phenotype BLUEs; μ is the overall mean; **g** is a vector of random effect of genotypes with $g \sim N(0, \boldsymbol{A}\sigma_g^2)$, where σ_g^2 is additive genetic variance and $\boldsymbol{A}$ is the realized additive relationship matrix calculated from $n \times m$ genotype matrix with n number of genotypes and m number of markers using *A.mat* function from rrBLUP package (Endelman and Jannink, 2012); and **e** is a vector of residuals that are identical and independently distributed with $e \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_e^2 is the residual variance and $\boldsymbol{I}$ is an identity matrix.

Estimation of genomic heritability

A re-parameterization of GBLUP model as explained in Janss, et al. (2012) was done to evaluate the impacts of population structure on genomic heritability of the traits. The reparameterized model can be written as:

$$y = 1\mu + U\alpha + e \tag{3}$$

In the above equation, **U** is an $n \times (n-1)$ matrix of the eigenvectors obtained from eigenvalue decomposition of additive relationship matrix (***A***) with $\mathbf{U}_i$ the column i ($i = 1, 2, \ldots, n-1$) of **U** representing the principal component loads; α is an $(n-1) \times 1$ vector of random effects with normal distribution $N(0, \mathbf{D}\sigma_g^2)$ where **D** is an $(n-1) \times (n-1)$ diagonal matrix with each diagonal element representing eigenvalues of ***A*** corresponding to that particular column. The model (3) with principal components as random variables generates the same marker distribution as model (2), and allows for separation of total genetic variance σ_g^2 into across-subpopulation genetic variance σ_{gA}^2 due to population structure, and within-subpopulation genetic variance σ_{gW}^2. This partitioning of total genetic variance allowed for estimation of within (h_{gA}^2) and across-subpopulation (h_{gW}^2) genomic heritabilities which were calculated as:

$$h_{gA}^2 = \frac{\frac{1}{n}\sum_{i=1}^{d}\alpha_i^2}{\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2 + \sigma_e^2} \tag{4.1}$$

and

$$h_{gW}^2 = \frac{\frac{1}{n}\sum_{i=d+1}^{n}\alpha_i^2}{\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2 + \sigma_e^2} \tag{4.2}$$

where d is largest eigenvectors in the population with n individuals used to account for population substructure that result in artifact variation arising due to population admixture, d was calculated using the *eigen* function on relationship matrix, ***A***. The posterior values for σ_{gA}^2, σ_{gW}^2, σ_g^2, and σ_e^2 were estimated by Markov Chain Monte Carlo (MCMC) using a Gibbs sampler as proposed by de los Campos, et al. (2010) and

Janss, et al. (2012) for each trait using phenotypic and genotypic data for all individuals in our panel. A total of 37,000 MCMC iterations were run with first 2000 iterations discarded for burn-in. The posterior means for within and across-subpopulation heritabilities were calculated form the estimated variance components.

Cross validation and prediction accuracy

Cross validation using stratified sampling (CV1)

A common cross validation approach using five-folds obtained by stratified sampling was done for CV method 1 (CV1). In stratified sampling, the individuals were proportionally sampled from each sorghum race to form cross-validation folds that have population structure similar to that of the whole population. As illustrated in Figure 1A, individuals within each race were randomly partitioned into five mutually exclusive groups (W_1, W_2, W_3, W_4, and W_5) with similar sample sizes resulting in five proportionally divided datasets (one per race). Then, five subsets (S_1, S_2, S_3, S_4, and S_5) were constructed such that each subset contained one of the partitions from each race (Figure 1A). During cross validation, each subset was treated as a fold, and four of the folds were assigned to the training set and the genetic values were predicted for the remaining fold. This process was repeated until every single fold and individuals were predicted only once, and the predicted genetic values for all individuals were stored. Prediction accuracy was calculated as correlation between predicted genetic values and observed phenotypic values of all individuals in the population for each cross validation run. A similar approach has previously been applied to study the effect of population structure in prediction results (Albrecht, et al., 2011, Guo, et al., 2014). Since all cross-validation folds are proportionally sampled from structured subpopulations, the training and validation sets used in prediction have similar racial structure. The accuracy from CV1 method was decomposed into covariances resulting from conditional expectations due to racial structure. The decomposition of covariance was calculated as described in Sorensen and Gianola (2007):

$$Cov(x,y) = E_{race}[Cov(x,y \mid race)] + Cov_{race}[E(x \mid race), E(y \mid race)] \quad (5)$$

where, x and y are predicted and observed values, respectively; E_{race} is expectation over races of the covariances within race; and Cov_{race} is covariance across races of the expectation within race. A multi-

response model with unstructured variances was fitted with scaled values of *x* and *y* as response variables, and race as a random variable in the model using the *MCMCglmm* function in the R package MCMCglmm (Hadfield, 2010). A total of 13,000 iterations were done with 3,000 burn-ins, and posterior mean was calculated from a total of 1000 estimates were recorded using a thinning interval of 10.

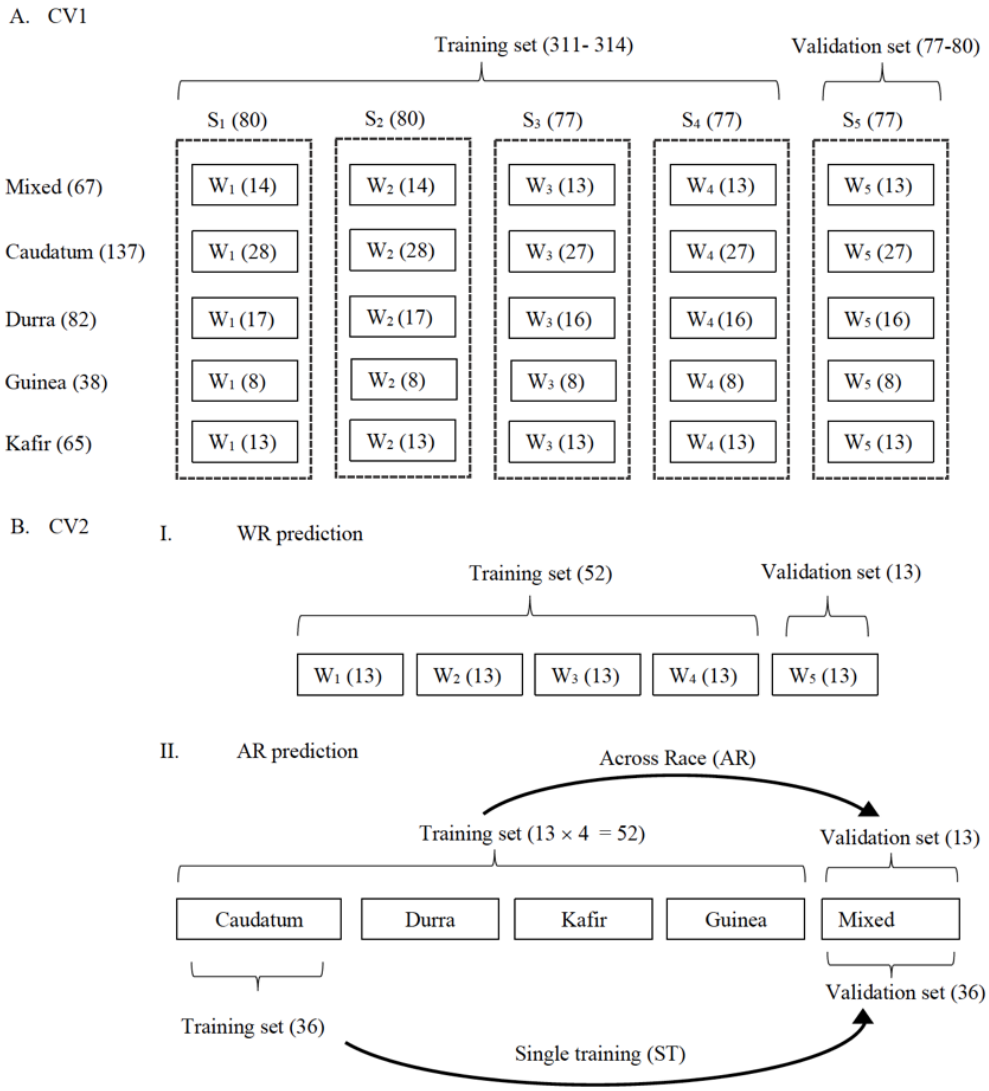


Figure 1. Examples for cross-validation approaches implemented in the sorghum diversity panel. A) CV, individuals in each race were proportionally divided into five datasets (W_1,…,W_5) and cross validation fold (S_1,…,S_5) was created as shown by rectangular boxes with broken lines ; B) CV2: within race (WR, I) and across race (AR, II) cross validation method for mixed race. A variation of AR prediction, single race training (SRT), was also implemented where a single race was used for training instead of all four races. In parentheses are the number of individuals used in prediction.

Across and within race cross validations (CV2)

While the CV1 method simulates similar population structure across both training and validation populations, breeding populations are often derived from genetically distinct pedigrees with dissimilar population structure. In order to understand how the GBLUP model for grain yield related traits in sorghum is affected by intrinsic racial structure, we designed a second cross validation experiment, CV2. In this approach, we ran predictions either by dividing individuals from a single race into training and validation folds, or by using individuals from certain race/s as a training population to predict genetic values of individuals from unrelated race/s (Fig 1B). Similar strategies have previously been reported for within and across group genomic prediction for diversity panels in maize and rice (Guo, et al., 2014), and for breeding population in wheat (Norman, et al., 2018).

The first CV2 method, within race (WR) prediction, was done by randomly dividing individuals within a single race into five proportional folds (Fig 1B). The five-folds are used for five-fold cross validation, the predicted values for individuals in each fold was stored and subsequently a single *r* was calculated for each cross validation run, as previously described for CV1 method. The five folds in this method are derived from the five mutually exclusive datasets (W_1, W_2, W_3, W_4, and W_5) that were used in CV1 for each race; however, in this method, four of these datasets/folds collectively formed the training set and the remaining dataset/fold was used as validation set. For each run, the predicted values were stored until each fold was predicted once, then a single *r* for calculated as correlation between predicted and observed phenotypic values for the given race.

In the second CV2 method, across race (AR) prediction was conducted using four of the races as training population and the fifth race as a validation race (Fig 1B). Unlike in CV1, AR doesn't have a uniform population structure across the folds and the individuals in the training and validation populations are from genetically distinct racial clusters. In order to maintain the same training population size between AR and WR predictions, we sampled proportional amount of individuals from each race to makeup the total cross-validation sample size equal to the sum total of individuals within the validation race. For example as shown in Figure 1B, a total of 13 individuals from each of the four races kafir, caudatum, durra, and guinea were sampled as training population ($n = 13 \times 4$) and breeding values were estimated for a random sample of 13

individuals from the mixed race. Subsequently, *r* for the mixed race was calculated as correlation for observed phenotypic values and predicted genetic values for those 13 randomly sampled individuals from the mixed race. Similarly, *r* for AR prediction of each race was calculated in similar fashion with a total of 13, 16, 27, and seven individuals sampled from each race for prediction of the races kafir, durra, caudatum, and guinea, respectively. In addition to AR method we also ran a variation of across subpopulation prediction, which we call single race training (SRT) method, where a single fold of 36 individuals sampled from a single race was used as the training population to obtain predicted genetic values of individuals from all other races (Fig 1B). The prediction accuracies for the SRT method were calculated as correlations between predicted and observed values for pairwise combination of training and validation races. The objective of this method was to explore the predictive relationship between any two races for a given trait.

A total of 100 random replications were conducted for each cross-validation method, and estimates for mean *r* and standard deviations were calculated. Vectoral graphs used in the analysis of results were created using various plotting functions in R package ggplot2 (Wickham, 2016).

Results

Racial structure

We identified an optimum of five subpopulation cluster based on estimates of cross-validation error from admixture (Supplementary Figure 1). Admixture ancestry coefficients (Q) were used to assign individuals into subpopulations, individuals with coefficients >50% were assigned into that subpopulation. Four of the five subpopulation clusters, thus identified, were broadly congruent with original racial classification of the accessions based on morphological characteristics (Figure 2a, Supplementary Datafile). The remaining accessions contained mixed ancestry based on admixture components, and a large proportion of them belonged to intermediate or mixed races based on original morphological classification (Casa, et al., 2008). For ease, the subpopulation clusters are referred to as corresponding race and “mixed” race represents the cluster of accessions with mixed or intermediate ancestry. Our results from admixture were supported by the principal component and neighbor joining analyses (Figure 2). In the neighbor joining tree, *S.*

propinquum, an outgroup individual which is a diploid wild sorghum from southeast Asia, clustered together with accessions from the race guinea suggesting potentially earlier adaptation and divergence of guinea race compared to caudatum, durra, and kafir (Figure 2c).

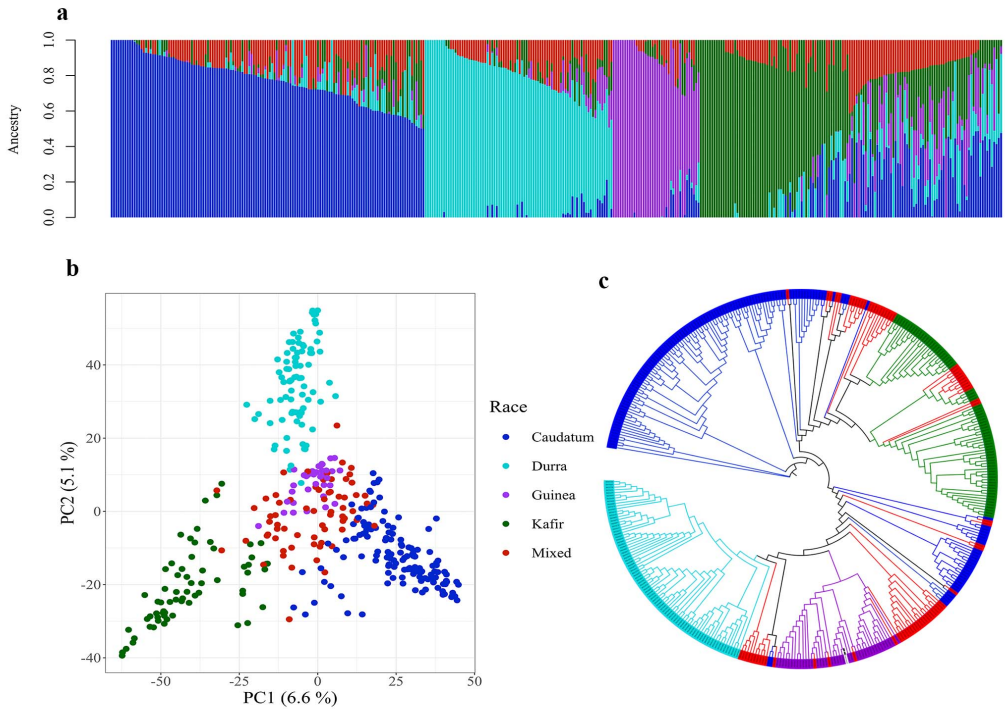


Figure 2. Population structure and clustering analysis of the sorghum diversity panel based on; a) ancestry coefficients for K=5 in admixture, b) principal component analysis of the first three PCs, and c) neighbor joining tree analysis. In parentheses, proportion of variation explained by the corresponding PC. Branches and labels in the tree and accessions in PCA are color coded by the race identified from population structure analysis using admixture. Branch represented by broken line in the guinea clade is wild sorghum *S. propinquum*.

Genetic diversity and linkage disequilibrium

Tajima's D and heterozygosity estimates suggest presence of strong genetic bottlenecks in our panel possibly from domestication, adaptation and artificial selection. The average distance over which LD decayed to half of its maximum value was ~20 kb in our panel, which is consistent with previous observations (Hamblin, et al., 2004, Mace, et al., 2013). Whereas, the average distance for LD decay to reach background levels ($r^2 < 0.1$) was around 100 kb, similar to previous observations of Morris, et al. (2013) in global diversity panels. We observed variability in the level of genetic diversity and LD among sorghum races (Table 1).

Average expected heterozygosity for all races were significantly different (p-value < 0.01) from each other. The presence of extensive LD, lower genetic diversity and Tajima's D values within kafir suggests the presence of a stronger genetic bottleneck within this race compared to others (Table 1). Among the five racial types, the mixed race had highest average nucleotide diversity (3.5×10^{-5}) and lowest LD. The race guinea had the highest average expected heterozygosity (0.5), whereas the average distance to LD decay for guinea were higher than all races except kafir. Genetic diversity and LD for durra were similar to that of the whole panel. Although LD and nucleotide diversity estimates of caudatum were comparable to that of guinea, heterozygosity in caudatum was about 30% of guinea (Table 1). We calculated the Euclidean distance between the centroids of five PCs and also estimated F_{st} for different races and found that kafir and mixed were genetically the most and least distant race, respectively, whereas the other three races (caudatum, durra, and guinea) seemed to be roughly equidistant from each other (Supplementary Table S1). These results show consistency with the timeline of diversification of these races, as kafir is probably the most recent and mixed race has some of the most primitive accessions of intermediate and bicolor race (Deu, et al., 2006, Doggett, 1988, Kimber, et al., 2013).

Table 1. Summary statistics of whole genome estimates for genetic diversity and LD.

	Whole panel	Mixed	Kafir	Durra	Caudatum	Guinea
Number of accessions	389	67	65	82	137	38
Average heterozygosity[a]	0.17	0.19	0.12	0.18	0.14	0.5
Nucleotide diversity (10^{-5})	3.23	3.48	2.32	3.07	3.01	3.12
Tajima's D	-0.63	-0.93	-1.2	-0.88	-1.01	-0.92
Average r^2	0.09	0.1	0.21	0.1	0.11	0.16
LD decay distance (bp)	20491	14625	145252	19870	32076	39712

[a]Nei's unbiased estimator of gene diversity (Nei, 1987)

Phenotypic variation and correlation

The differences between the population means of at least some of the races were significantly different (p-value < 10^{-5}) for all traits except FLH and PH. Phenotypic mean and standard deviation for individual races are listed in Supplementary Table S2. Variation in phenotypic distribution and correlation between traits was observed across all traits (Figure 3). Grain yield was positively correlated with both GN and GW while the two yield components (GN and GW) were slightly negatively correlated to each other.

While BL showed a significantly negative correlation with grain yield traits, remaining plant and inflorescence architecture traits (PH, FLH and PL) didn't exhibit any significant correlation to grain yield traits. Grain yield, GN, PH, FLH were all significantly positively correlated with DTA, whereas PL and GW showed significantly negative correlation with DTA.

Prediction accuracy and racial structure

Mean prediction accuracy of a trait is known to be directly related to its heritability (Combs and Bernardo, 2013). We were interested in the nature of relationship between CV1 prediction accuracy and total genomic heritability across all traits. Therefore, we ran correlation using mean estimates of CV1 accuracy from all traits to their respective genomic heritabilities and observed a strong correlation (0.63) between the two. For the CV1 method, the highest and lowest *r* were 0.69 for GW and 0.52 for GN and DTA, respectively. Among the traits studied, PH and GN had the highest and lowest genomic heritabilities, respectively (Figure 4a). Despite low heritability, GY had a mean *r* of 0.57 for the CV1 method and DTA had lowest *r* (0.52) despite moderate-high genomic heritability (0.73). As expected, *r* from the CV1 method were always higher than *r* from CV2 prediction methods for all traits, which can be ascribed to the larger training population size and similar population structure between training and validation population in CV1 (Table 2).

Figure 4b shows decomposition of the CV1 accuracy into covariances resulting from conditional expectation of races. The scaled covariances E_{race} and Cov_{race} represent expectation due to race and covariances due to individuals within race, respectively. The mean covariances E_{race} and Cov_{race} were positively correlated with posterior means of across (0.61) and within race (0.81) genomic heritabilities, respectively. The estimates for covariances E_{race} and Cov_{race} were comparable to estimates of *r* for AR and WR prediction, respectively, except for height. The estimates of covariances and variance of predicted values for AR prediction method were smaller than in WR (Supplementary Table S3). Hence, the differences in mean *r* between the two methods weren't as pronounced as seen in rice and maize (Guo, et al., 2014).

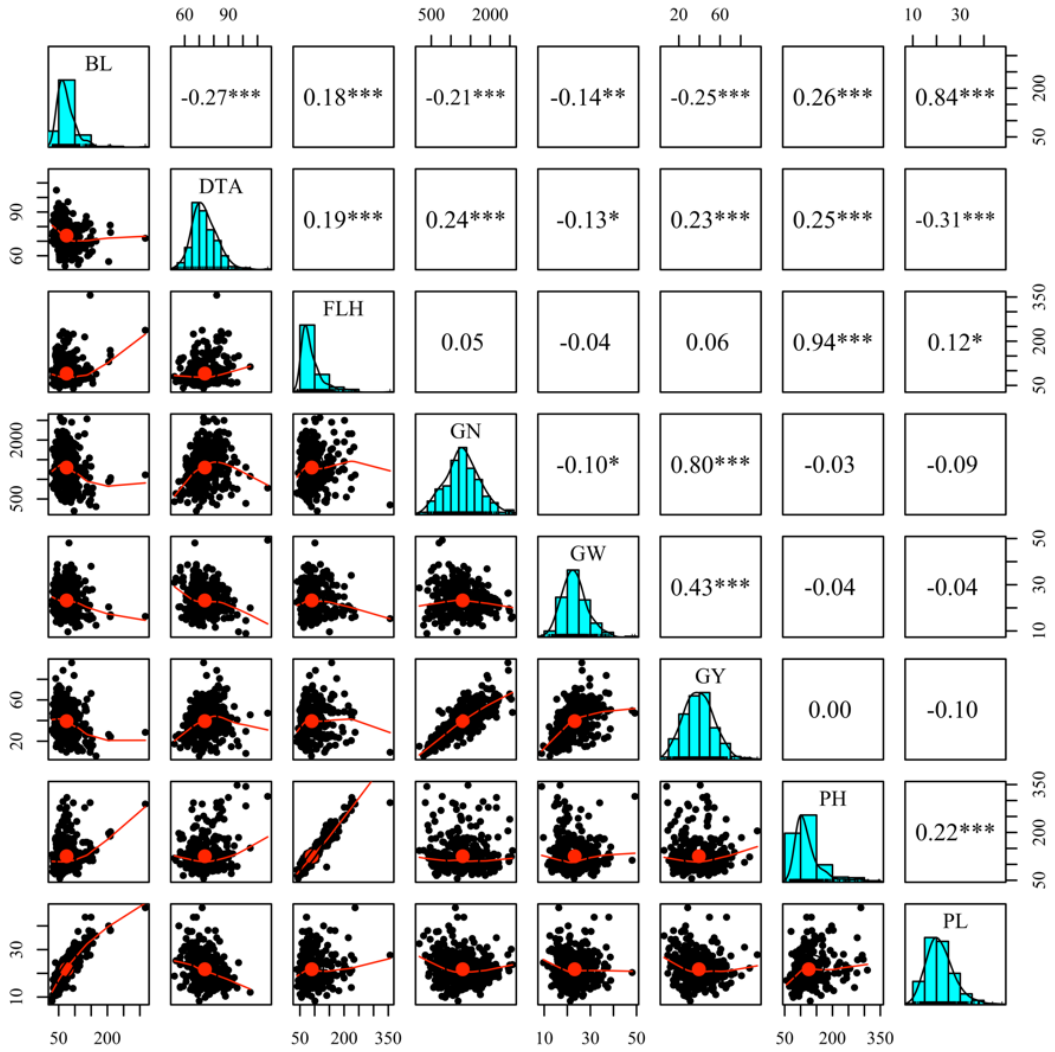


Figure 3. Distribution and pairwise correlations for adjusted phenotypic mean for all eight traits. Histograms for traits is displayed along the diagonal. Scatterplots with line of fit (red line) for all individuals in the diversity panel are to the left and below the diagonal. Pearson correlation coefficient between the traits shown above the diagonal and to the right. Significance level: *p < 0.05, **p < 0.01, **p < 0.001. BL: panicle branch length, DTA: days to anthesis, FLH: flag leaf height, GN: grain number per primary panicle, GW: thousand grain weight, GY: grain yield per primary panicle, PH: plant height, PL: panicle length.

Table 2. Mean prediction accuracy (*r*) of different cross validation methods for all traits studied. Values represent mean ± standard deviation. CV1: cross validation method-1, AR: across race, WR: within race, SRT: single race training.

Trait	CV1	WR	AR	SRT
Days to anthesis	0.52 ± 0.01	0.24 ± 0.23	0.12 ± 0.32	0.11 ± 0.23
Flag leaf height	0.58 ± 0.03	0.41 ± 0.23	0.34 ± 0.33	0.32 ± 0.21
Grain number/panicle	0.52 ± 0.02	0.25 ± 0.12	0.27 ± 0.30	0.26 ± 0.20
1000-grain weight	0.69 ± 0.02	0.61 ± 0.10	0.37 ± 0.27	0.43 ± 0.20
Grain yield/panicle	0.57 ± 0.02	0.36 ± 0.12	0.35 ± 0.30	0.38 ± 0.17
Plant height	0.63 ± 0.02	0.46 ± 0.15	0.45 ± 0.28	0.36 ± 0.18
Panicle length	0.65 ± 0.01	0.25 ± 0.31	0.12 ± 0.33	0.14 ± 0.21
Terminal branch length	0.67 ± 0.07	0.38 ± 0.24	0.20 ± 0.31	0.26 ± 0.19

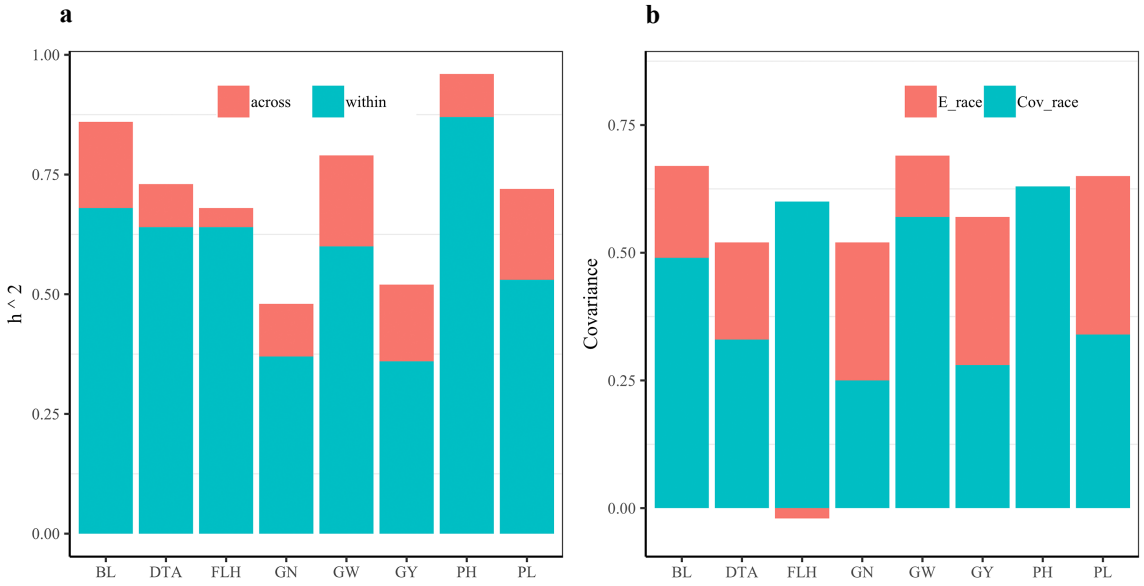


Figure 4. Posterior means of, **a**) within-subpopulation and across-subpopulation genomic heritabilities using first five principal components, **b**) scaled covariances due to condition expectation of race in CV1 prediction. *E_race* represents covariance due to race, and *Cov_race* represents covariance due to individuals within race. h^2, genomic heritability; PH, plant height; GW, thousand grain weight; FLH, flag leaf height; GN, grain number per panicle; GY, grain yield per primary panicle; DTA, days to anthesis; BL, primary branch length; PL, panicle length.

Among CV2 prediction methods, estimates of *r* for WR were higher than AR for BL, DTA, FLH, GW and PL, but the two methods had similar estimates for PH, GN and GY (Table 2). However, *r* for different CV2 methods varied depending on the combination of trait and race (Supplementary Table S4). Posterior means of within-subpopulation genomic heritabilities were moderately correlated (0.4) to mean *r*

from WR prediction. The average *r* for WR were higher than AR for caudatum and durra whereas the two methods had similar averages for guinea, kafir, and mixed (Supplementary Table S4). Plant height and FLH showed smaller difference in *r* between AR and WR prediction across all races, whereas GW showed consistently higher estimates of *r* for WR over AR for all races. Among all the traits, GW had highest mean *r* for WR (0.61) and SRT (0.43) prediction methods, while PH (0.45) had highest *r* for AR.

In SRT method, the traits that are heavily correlated to racial structure (DTA, PL, BL, GN and GY) were poorly predicted than PH, FLH and GW (Figure 5). In general, the races durra and caudatum resulted in poor prediction for SRT method when introduced in model as training or validation populations compared to mixed and guinea races (Figure 5). The two former races are thought to have diverged recently than the latter two (Kimber, et al., 2013). So we conducted an additional across race cross-validation using a random subset of 36 kafir accessions as validation population and a combination of 36 accessions from one or many remaining races as training population (Supplementary Figure S2). We started with 36 accessions from mixed race based on earlier divergence and best predictor of kafir in CV2 SRT prediction results. Mixed race by itself predicted as good as or better than most of combination which is expected due to the closer relationship and larger diversity of the race (Supplementary Figure S2). However, a combination of guinea and mixed performed the best for plant height and grain number. Grain yield showed increase in accuracy with combination of various races except in the case of MD (mixed-durra). While mixed race is more closely related to all other races, consistently better performance of guinea and kafir as validation population in SRT could be due the amount of shared (versus population-specific) allelic variation in these groups. We observed higher proportions of intermediate frequency minor alleles (0.1 to 0.4) in mixed and guinea than in caudatum, durra and kafir (Supplementary Figure S3). We identified private alleles in each races as the minor alleles that were present in a particular race and were absent in all other races. Within the mixed race group, there were 3,378 private polymorphisms that were not present in any of the other races, although on average these were only present at low frequencies within the population mean (MAF = 0.03). Caudatum had fewer private polymorphisms (1,843), with a mean MAF of 0.04. Durra, on the other hand, had the highest number of private SNPs with 3,969 and the highest mean MAF for these sites (0.07). The races kafir and guinea had no private alleles.

In order to assess the effect of training population size in *r* for AR and WR method, we ran cross validations using accessions in race caudatum using training population sizes of 28, 52, 74, 96, and 110. We ran cross validations only for PH, GN, GW, and GY because they have varying trait genetic architecture, PH and GW have relatively high genomic heritability, whereas GY and GN have relatively low heritability. So we reasoned using these four traits will be sufficient to deduce necessary information about the role of training population size, while keeping the analysis relatively simple. We observed that while increasing training population size showed consistent increase in *r* for WR prediction of all four traits, increasing training population size did not always lead to increased *r* in AR prediction (Figure 6). Mean *r* for WR was always higher than AR for GW across all training population sizes, whereas they were similar for PH among the two methods. For GN and GY, *r* was higher for AR prediction when training population size was 28 individuals and similar when training population size was 52. At larger training population size, WR prediction method had larger *r* than AR prediction.

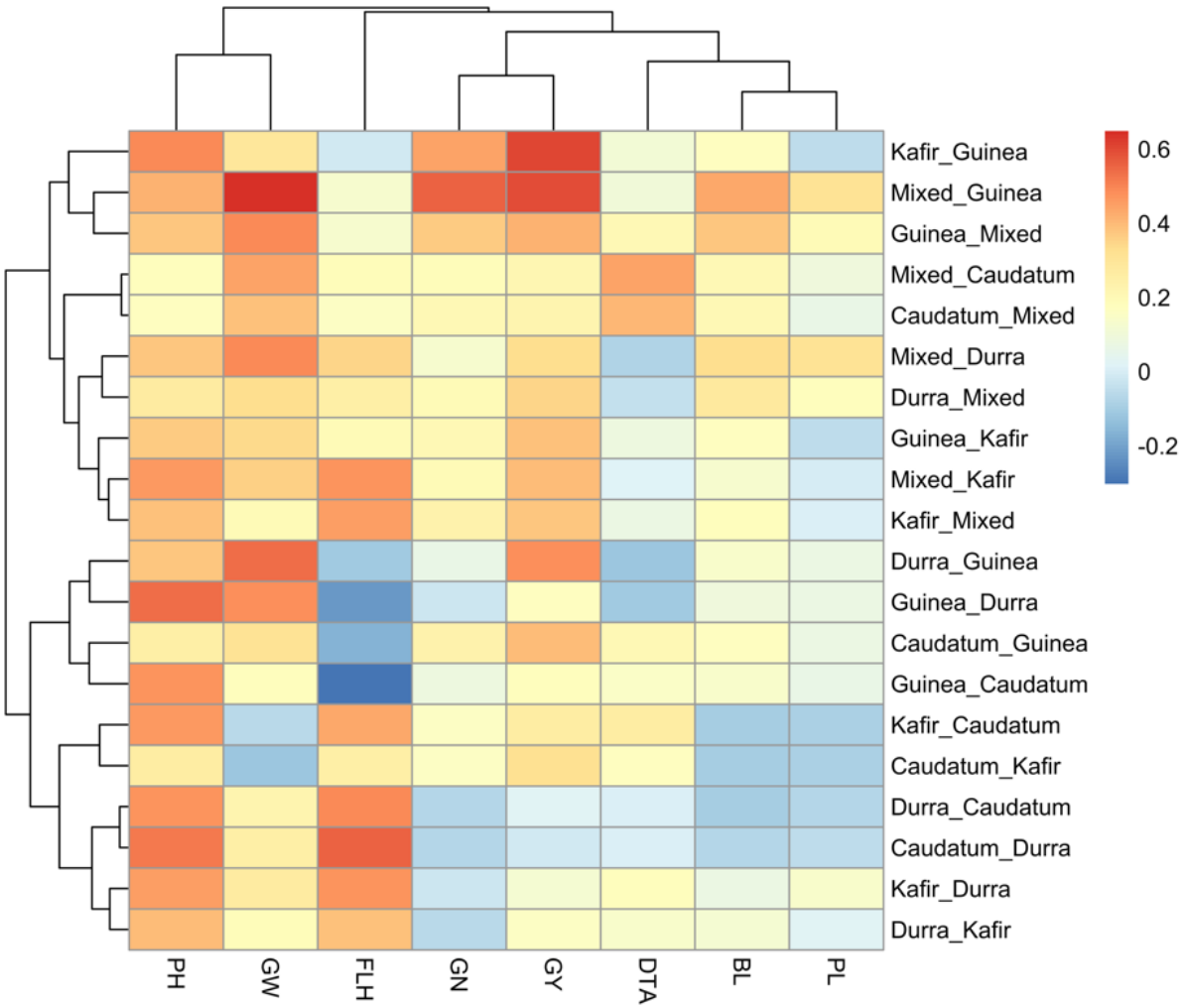


Figure 5. Heatmap showing mean prediction accuracies (*r*) from pairwise single race (SRT) prediction in CV2 prediction method. The races to the right of the heatmap represent training race followed by validation race. Tree cluster to the top and left is based on hierarchical clustering of the values from column and rows, respectively. PH, plant height; GW, thousand grain weight; FLH, flag leaf height; GN, grain number per panicle; GY, grain yield per primary panicle; DTA, days to anthesis; BL, primary branch length; PL, panicle length.

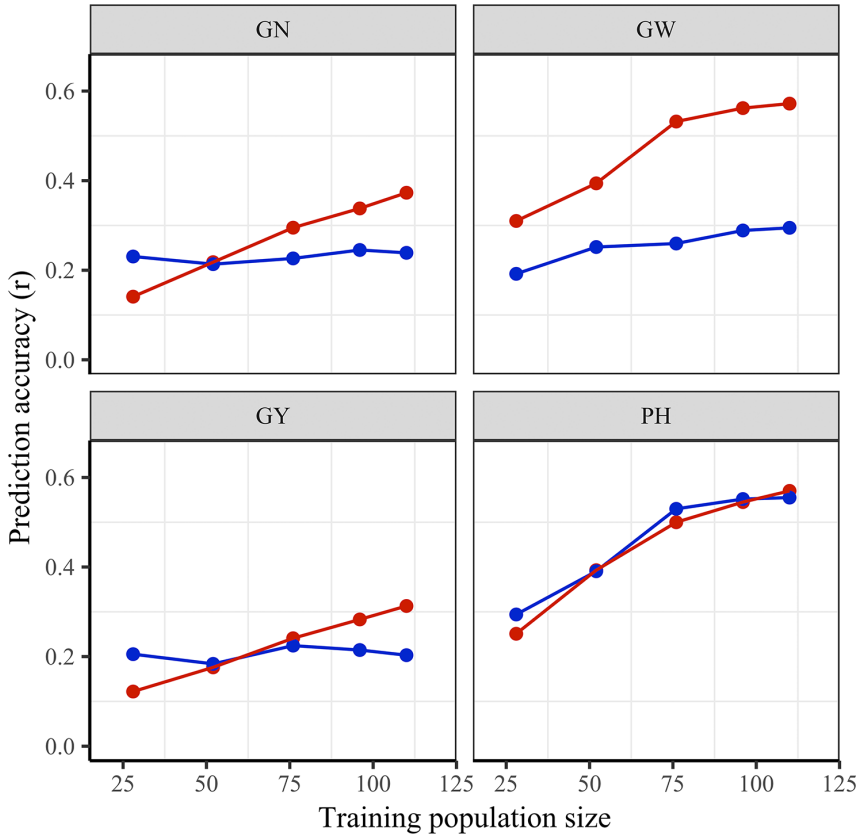


Figure 6. Mean prediction accuracies from across race (AR) and within race (WR) prediction methods for different training population sizes in caudatum. Colors represent cross validation methods, blue = AR, red = WR. GN = grain number, GW = grain weight, GY = grain yield, PH = plant height.

Discussion

Genetic differentiation and racial structure

The results from the population structure analysis support five subpopulation clusters in our sorghum diversity panel. While the four subpopulations were congruent with the four most recent sorghum races, the bicolor race was not diverged enough to form one distinct group. This is consistent with previous observations from clustering analysis in populations consisting of diverse sorghum accessions (Brown, et al., 2011, Deu, et al., 2006, Wang, et al., 2013). The hypothesis of Harlan and Stemler (1976) that the guinea race was probably the earliest to have diverged from the early bicolor domesticates is supported by our neighbor joining analysis (Figure 2c). *Sorghum propinquum*, an outgroup diploid species from southeast Asia,

clustering together with the guinea clade suggests guinea potentially diverged from the early bicolor prior to the divergence of kafir, durra and caudatum. Furthermore, we also see a large proportion of shared alleles and higher allelic diversity among guinea than the evolutionarily recent races, which, however, could also be due to higher rate of gene flow.

Genetic diversity and linkage disequilibrium

Our diversity panel is representative of much of the genetic, phenotypic and geographic diversity of the global sorghum germplasm, and therefore it is an excellent resource for genetic dissection of agronomically important traits and adaptation. Patterns of LD, Tajima's D and genetic diversity from the whole population and within different races suggest strong genetic bottlenecks in our population as a result of domestication, adaptation, and artificial selection. The rate of LD decay, Tajima's D and genetic diversity in our population were comparable to estimates from previous studies (Hamblin, et al., 2004, Mace, et al., 2013, Wang, et al., 2013). Average distance to half decay of LD was similar to that reported by Mace, et al. (2013) for improved inbreds, and lower than the previous estimates of Morris, et al. (2013). Wang, et al. (2013) using 242 diverse accessions from a sorghum mini core collection and 13,390 SNPs, found that LD decayed to background levels between 10 and 30 kb for all but chromosome 2. Average r^2 for different sorghum races in our study were consistent with estimates of Wang et al. (2013). Genetically least diverse racial types (kafir and caudatum) showed higher extents of LD than the races with higher diversity, with the exception of guinea. Previous studies involving separate diversity panels have also reported reduced diversity in kafir (Deu et al. 2006; Casa et al. 2008). The extensive LD and lower heterozygosity in kafir and caudatum might be the result of limited cross-pollination and geographical isolation, whereas genetic drift and smaller sample size could have contributed to slower rate of LD decay in guinea. Bouchet, et al. (2012) also observed similar outcomes for LD and genetic diversity among sorghum races. The lower abundance of private alleles among races kafir and guinea than caudatum and durra was observed by Bouchet, et al. (2012), which is consistent with lack of private alleles for guinea and kafir in our population. Despite being the oldest and most heterozygous race

guinea didn't possess any private allele, which could be a result of small sample size of guinea in our population.

Genomic prediction and racial structure

Heritabilities and CV1 prediction accuracy for grain yield and plant height in our study were similar to previous studies in sorghum (Fernandes, et al., 2018, Hunt, et al., 2018, Yu, et al., 2016). Similar heritabilities and *r* for flowering time, panicle length, plant height and grain number have previously been reported in a rice diversity panel (Guo, et al., 2014) . The overwhelming contribution of within-subpopulation genomic heritabilities towards the total genomic heritabilities of traits is comparable to previous observation in rice (Guo, et al., 2014). Grain yield was predicted with consistently higher accuracy than grain number across all cross validation methods despite similar heritabilities of the two traits. While both GN and GY are highly correlated complex traits controlled by a large number of small effect loci, higher accuracy of GY over GN could have resulted from strong positive correlation (0.43) of GY to GW, which has the highest *r*. Traits controlled by large number of small effect loci are predicted with higher accuracy when higher allelic diversity exists in the training population as compared to traits governed by few relatively large effect loci (Norman, et al., 2018). With smaller training sizes for GN and GY the breadth of genetic diversity from all races might have led to boost in *r* for AR over WR, but as training size increased the genomic relatedness in WR appeared to outweigh the effect of genetic diversity, which resulted in stronger positive relationship between *r* and training size in WR (Figure 6). Since the range of training size in our study is small, genetic diversity did not increase substantially with increasing training size resulting in lack of linear relationship between training size and *r* in AR for GN and GY.

Population structure resulting from domestication and diversifying selection leads to varying levels of genetic relatedness among individuals within and between subpopulations. The accuracy with which breeding values are estimated is affected by stratification in the population, and the effects are more pronounced when the genetic architecture of the predicted trait is directly associated with population structure (Isidro, et al., 2015, Windhausen, et al., 2012). Previous studies have shown that when population structure

in training and testing populations is similar, it can contribute positively towards prediction accuracy (Bastiaansen, et al., 2012, Crossa, et al., 2014, Guo, et al., 2014). But when cross validation strategies that constrained population structure were implemented, it resulted in decline of *r* due to the weakened genetic relationship among individuals in training and testing population (Guo, et al., 2014, Ly, et al., 2013, Lyra, et al., 2018, Norman, et al., 2018).

In order to understand the contribution of racial structure in prediction accuracy of stratified sampling method, we decomposed the CV1 accuracy into expectation over races (E_{race}) and covariance due to individuals within race (Cov_{race}). Almost non-existent E_{race} covariances for the two height traits, FLH and PH, indicates race as a predictor contributed poorly towards prediction of height. On the other hand, race contributed relatively larger proportion of total covariance for grain yield components and panicle architecture traits. Since the racial structure of sorghum can be directly associated with panicle architecture and indirectly to the grain yield components, proportion of across race genomic heritability and covariance due to race were higher for those traits. We saw sharper decline in *r* for AR prediction compared to WR, especially for panicle architecture traits, which could be attributed to poor genomic relationship between training and validation population. Height traits, PH and FLH, that are less associated with racial structure showed smaller decline in *r* than panicle architecture and grain yield traits. The SRT prediction method also showed smaller *r* for pairwise prediction results for the panicle architecture traits than other traits. Yu, et al. (2016) have previously observed that race as a predictor explains higher variation in predicted values of biomass traits than actual phenotypic values in sorghum, suggesting that under the presence of similar racial structure in training and validation population the accuracy of genomic prediction might have been inflated as a result of overemphasis on racial differences (Brown, 2016). Our approach of decomposition of covariances into conditional expectations due to race could be utilized in dissection of impact of population structure in cross validation accuracy from stratified and random sampling methods in diverse as well as breeding populations.

Cross validation approaches similar to the one employed in this study have resulted in higher *r* for within population prediction than across population prediction in wheat (Norman, et al., 2018), rice and maize (Guo, et al., 2014). Although average *r* across all races in our study was higher for within population for most

of the traits, the variation in *r* for individual race and trait combination shows interaction between population structure and trait genetic architecture. Higher *r* for WR over AR among the races caudatum and durra could be because of higher proportion of private alleles and smaller proportion of intermediate frequency minor alleles. This could be the reason for smaller difference between average *r* for AR and WR among guinea, kafir and mixed, as these races show lack of private alleles and/or higher genetic diversity. Furthermore, the results from clustering analysis have shown that the mixed race has closest genetic relationship to the rest of the four races. Our SRT prediction which was a good measure of pairwise predictive relationship between two races also shows that kafir, guinea and mixed have better predictive relationship to each other than to caudatum and durra. This was further supported by our cross-validation using various combination of races to predict kafir, evolutionarily the most recent race (Supplementary Figure S2). Genetic diversity and divergence seems to have an important impact in prediction accuracy, which needs to be an important consideration during training population design for diverse germplasm evaluation.

Potential applications for sorghum breeding

Genomic prediction was first introduced roughly two decades ago and has been applied in plant breeding for over a decade (Bernardo and Yu, 2007, Meuwissen, et al., 2001). However, studies investigating prospects and applications of genomic prediction in sorghum are limited. A few studies have been reported for biomass traits in diversity panels (Yu et al. 2016; Fernandes et al. 2017), grain yield in pedigreed male inbred lines (Hunt et al. 2018), and a simulation study investigating prospects in a small sorghum breeding program (Muleta et al. 2019). While clearly defined heterotic pools do not exist in sorghum as they do in maize, races have long been exploited by sorghum breeding programs for hybrid production. If we are to exploit the vast genetic diversity of races in sorghum breeding, we need a more comprehensive understanding of how racial structure impacts prediction accuracy for economically and agronomically important traits.

Our study suggests maintaining a genetically diverse training population that includes a mixed/intermediate race might boost prediction accuracy when training population size is constrained. This strategy might be beneficial for young and small breeding programs where breeders have limited resource to

construct individual training population for different breeding populations (Muleta, et al., 2019). Furthermore, new phenotypic data from diverse lines when added into the training population can allow for maintenance and increase in frequency of advantageous minor alleles in the gene pool. Guinea had the highest mean prediction accuracy for grain yield and grain weight irrespective of prediction method, suggesting a genetically diverse training population is likely to predict the yield potential of best performing guinea more accurately than individuals from any other race. So breeding programs in West Africa, where guinea sorghum is widely grown, could utilize genotypic and phenotypic data from all racial types in training population design for genomic prediction of guinea varieties. Our results from the SRT method showed that moderate prediction accuracy can be gained even by using a single completely unrelated race as a training population for grain yield components and height. Historically, sorghum breeding programs in the US have heavily relied on kafir and caudatum types while genetic diversity from other races are underutilized. While breeding programs with plentiful resources could gain higher selection accuracy by simply increasing training population size and designing several independent training populations, the utilization of interracial diversity in genomic prediction could help in introducing novel sources of variation for diseases and pest resistance as well as genetic variation for increasing yield potential in the long run. Similarly, another way to increase selection accuracy and genetic gain of complex traits is through utilization of trait-assisted and indirect genomic selection when highly heritable and correlated secondary traits are available (Fernandes et al. 2017). For example, durra accessions in our results show a within race prediction accuracy of 0.33 and 0.43 for grain number and grain yield but an accuracy of 0.81 and 0.79 for branch length and panicle length, respectively. Panicle architecture could be used in indirect or trait-assisted genomic prediction for grain yield by breeding programs dominated by durra type sorghum varieties. In addition to utilization of within and across group genetic variances, optimization algorithms could also help in efficient design of training population for diversity panels and breeding populations (Akdemir, et al., 2015, Isidro, et al., 2015).

For effective use of crop diversity in sorghum breeding, a breeder might want to work with best representatives from all races rather than opting for the best lines of some races (Brown, 2016). In practical applications, prediction accuracy of traits that are affected by population structure can be increased by using genetically distant subpopulations as parental lines (Guo, et al., 2014). Our results can be useful in such an

effort because understanding how individuals of certain race respond to models trained using unrelated races can provide insights into how overall genetic diversity can be deployed in prediction of different racial types. For example, the prediction results from our SRT method shows guinea or mixed race with 37 individuals predicted GN and GY in kafir with higher accuracy than from using 52 kafir accessions (Figure 3, Supplementary Table S3). This kind of empirical evidences of predictive relationship can help in identifying trait-specific and race-specific training design for genomic prediction highlighting the need for more explorative and empirical case studies in natural and breeding populations.

Conclusion

Similar population structure between training and testing population can have positive impact on accuracy of genomic prediction. However, inflation in prediction accuracy could be an outcome of genomic prediction models overemphasizing racial differences. Prediction accuracy among races with higher proportion of allelic diversity and/or shared alleles is boosted by training population with higher genetic diversity despite poor genomic relationship, whereas genomic relationship outweighed genetic diversity among races with limited diversity and/or presence of unique polymorphisms. Therefore, training population design for a historically diverse and structured population in sorghum requires careful consideration of genetic structure of the testing population. While the sorghum association panel (SAP) was not intended for genomic prediction, the breadth of genetic and phenotypic diversity in this panel can allow for its application as training population for estimation of breeding values of diverse gene bank accessions. To that objective, including more guinea and bicolor accessions to the panel would be beneficial because our results show that accessions in these races boosts prediction accuracy as training and testing population.

Acknowledgements

We would like to thank the endowment fund for the Robert and Lois Coker Trustees Chair of Genetics, Wade Stackhouse fellowship, and Clemson University's Public Service and Agriculture agency for their support

for our research and training. We are grateful to Jianming Yu, Jim Holland, Jean-Luc Jannink and our reviewers for their helpful comments and suggestions.

Supplemental material

Supplementary file consists of three supplementary figures and four supplementary tables. Supplementary file is in Appendix A.

Conflict of interest

The authors declare no conflict of interest.

References

Akdemir, D., J.I. Sanchez and J.-L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47: 38.

Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer and C.C. Schon. 2011. Genome-based prediction of testcross values in maize. Theor Appl Genet 123: 339-350. doi:10.1007/s00122-011-1587-7.

Alexander, D.H., J. Novembre and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655-1664. doi:10.1101/gr.094052.109.

Bastiaansen, J.W., A. Coster, M.P. Calus, J.A. van Arendonk and H. Bovenhuis. 2012. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol 44: 3. doi:10.1186/1297-9686-44-3.

Bates, D., M. Mächler, B. Bolker and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67: 48. doi:10.18637/jss.v067.i01.

Bernardo, R. and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science 47: 1082-1090.

Bouchet, S., D. Pot, M. Deu, J.F. Rami, C. Billot, X. Perrier, R. Rivallan, L. Gardes, L. Xia, P. Wenzl, A. Kilian and J.C. Glaszmann. 2012. Genetic structure, linkage disequilibrium and signature of selection in Sorghum: lessons from physically anchored DArT markers. PLoS One 7: e33470. doi:10.1371/journal.pone.0033470.

Boyles, R.E., E.A. Cooper, M.T. Myers, Z. Brenton, B.L. Rauh, G.P. Morris and S. Kresovich. 2016. Genome-Wide Association Studies of Grain Yield Components in Diverse Sorghum Germplasm. Plant Genome 9. doi:10.3835/plantgenome2015.09.0091.

Boyles, R.E., B.K. Pfeiffer, E.A. Cooper, B.L. Rauh, K.J. Zielinski, M.T. Myers, Z. Brenton, W.L. Rooney and S. Kresovich. 2017. Genetic dissection of sorghum grain quality traits using diverse and segregating populations. Theor Appl Genet 130: 697-716. doi:10.1007/s00122-016-2844-6.

Brown, P.J. 2016. Plant breeding: Effective use of genetic diversity. Nat Plants 2: 16154. doi:10.1038/nplants.2016.154.

Brown, P.J., P.E. Klein, E. Bortiri, C.B. Acharya, W.L. Rooney and S. Kresovich. 2006. Inheritance of inflorescence architecture in sorghum. Theor Appl Genet 113: 931-942. doi:10.1007/s00122-006-0352-9.

Brown, P.J., S. Myles and S. Kresovich. 2011. Genetic support for phenotype-based racial classification in sorghum. Crop Science 51: 224-230.

Casa, A.M., G. Pressoir, P.J. Brown, S.E. Mitchell, W.L. Rooney, M.R. Tuinstra, C.D. Franks and S. Kresovich. 2008. Community resources and strategies for association mapping in sorghum. Crop Science 48: 30-40.

Combs, E. and R. Bernardo. 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Ceron-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett and K. Mathews. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112: 48-60. doi:10.1038/hdy.2013.16.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156-2158. doi:10.1093/bioinformatics/btr330.

de los Campos, G., D. Gianola, G.J. Rosa, K.A. Weigel and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92: 295-308.

de los Campos, G. and D. Sorensen. 2014. On the genomic analysis of data from structured populations. J Anim Breed Genet 131: 163-164. doi:10.1111/jbg.12091.

Deu, M., F. Rattunde and J. Chantereau. 2006. A global view of genetic diversity in cultivated sorghums using a core collection. Genome 49: 168-180. doi:10.1139/g05-092.

Doggett, H. 1988. Sorghum. Longman Scientific and Technical, New York, N.Y.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4: 250-255. doi:10.3835/plantgenome2011.08.0024.

Endelman, J.B. and J.L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3 2: 1405-1413. doi:10.1534/g3.112.004259.

Fernandes, S.B., K.O.G. Dias, D.F. Ferreira and P.J. Brown. 2018. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet 131: 747-755. doi:10.1007/s00122-017-3033-y.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun and E.S. Buckler. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9: e90346. doi:10.1371/journal.pone.0090346.

Guo, Z., D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang and G. Gay. 2014. The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127: 749-762. doi:10.1007/s00122-013-2255-x.

Gusnanto, A., C.C. Taylor, I. Nafisah, H.M. Wood, P. Rabbitts and S. Berri. 2014. Estimating optimal window size for analysis of low-coverage next-generation sequence data. Bioinformatics 30: 1823-1829. doi:10.1093/bioinformatics/btu123.

Hadfield, J.D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33: 1-22.

Hamblin, M.T., S.E. Mitchell, G.M. White, J. Gallego, R. Kukatla, R.A. Wing, A.H. Paterson and S. Kresovich. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. Genetics 167: 471-483.

Harlan, J.R. and J.M.J. de Wet. 1972. A Simplified Classification of Cultivated Sorghum1. Crop Science 12: 172-176. doi:10.2135/cropsci1972.0011183X001200020005x.

Harlan, J.R. and A. Stemler. 1976. The races of sorghum in Africa. In: J. R. Harlan, J. M. J. de Wet and A. Stemler, editors, Origins of African plant domestication. Mouton Publishers, Paris. p. 465-478.

Hill, W.G. and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33: 54-78.

Hunt, C.H., F.A. van Eeuwijk, E.S. Mace, B.J. Hayes and D.R. Jordan. 2018. Development of Genomic Prediction in Sorghum. Crop Science 58: 690-700. doi:10.2135/cropsci2017.08.0469.

Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot and M.E. Sorrells. 2015. Training set optimization under population structure in genomic selection. Theor Appl Genet 128: 145-158. doi:10.1007/s00122-014-2418-4.

Janss, L., G. de Los Campos, N. Sheehan and D. Sorensen. 2012. Inferences from genomic models in stratified populations. Genetics 192: 693-704. doi:10.1534/genetics.112.141143.

Kimber, C.T., J.A. Dahlberg and S. Kresovich. 2013. The gene pool of Sorghum bicolor and its improvement. Genomics of the Saccharinae. Springer. p. 23-41.

Letunic, I. and P. Bork. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44: W242-245. doi:10.1093/nar/gkw290.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G. Dixon, P. Kulakow and J.-L. Jannink. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Science 53: 1312-1325.

Lyra, D.H., Í.S.C. Granato, P.P.P. Morais, F.C. Alves, A.R.M. dos Santos, X. Yu, T. Guo, J. Yu and R. Fritsche-Neto. 2018. Controlling population structure in the genomic prediction of tropical maize hybrids. Molecular Breeding 38: 126. doi:10.1007/s11032-018-0882-2.

Mace, E.S., S. Tai, E.K. Gilding, Y. Li, P.J. Prentis, L. Bian, B.C. Campbell, W. Hu, D.J. Innes, X. Han, A. Cruickshank, C. Dai, C. Frere, H. Zhang, C.H. Hunt, X. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Lin, I.D.

Godwin, D.R. Jordan and J. Wang. 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4: 2320. doi:10.1038/ncomms3320.

McCouch, S., G.J. Baute, J. Bradeen, P. Bramel, P.K. Bretting, E. Buckler, J.M. Burke, D. Charest, S. Cloutier, G. Cole, H. Dempewolf, M. Dingkuhn, C. Feuillet, P. Gepts, D. Grattapaglia, L. Guarino, S. Jackson, S. Knapp, P. Langridge, A. Lawton-Rauh, Q. Lijua, C. Lusty, T. Michael, S. Myles, K. Naito, R.L. Nelson, R. Pontarollo, C.M. Richards, L. Rieseberg, J. Ross-Ibarra, S. Rounsley, R.S. Hamilton, U. Schurr, N. Stein, N. Tomooka, E. van der Knaap, D. van Tassel, J. Toll, J. Valls, R.K. Varshney, J. Ward, R. Waugh, P. Wenzl and D. Zamir. 2013. Agriculture: Feeding the future. Nature 499: 23-24. doi:10.1038/499023a.

Meuwissen, T.H., B.J. Hayes and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Morris, G.P., P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, H.D. Upadhyaya, O. Riera-Lizarazu, P.J. Brown, C.B. Acharya, S.E. Mitchell, J. Harriman, J.C. Glaubitz, E.S. Buckler and S. Kresovich. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci 110: 453-458. doi:10.1073/pnas.1215985110.

Muleta, K.T., G. Pressoir and G.P. Morris. 2019. Optimizing Genomic Selection for a Sorghum Breeding Program in Haiti: A Simulation Study. G3 9: 391-401. doi:10.1534/g3.118.200932.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Norman, A., J. Taylor, J. Edwards and H. Kuchel. 2018. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3 8: 2889-2899. doi:10.1534/g3.118.200311.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419-420. doi:10.1093/bioinformatics/btp696.

Price, A.L., N.A. Zaitlen, D. Reich and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459-463. doi:10.1038/nrg2813.

R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman and E.S.t. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci 98: 11479-11484. doi:10.1073/pnas.201394398.

Revelle, W.R. 2011. psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.1-2. Evanston, Illinois.

Sorensen, D. and D. Gianola. 2007. Likelihood, Bayesian, and MCMC methods in quantitative genetics.Springer. (Page 67).

Swarts, K., H. Li, J.R. Navarro, D. An, M. Romay, S. Hearne, C. Acharya, J. Glaubitz, S. Mitchell and R. Elshire. 2014. FSFHap (Full-Sib Family Haplotype Imputation) and FILLIN (Fast, Inbred Line Library ImputatioN) optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7: 1-12.

Technow, F., C. Riedelsheimer, T.A. Schrag and A.E. Melchinger. 2012. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125: 1181-1194. doi:10.1007/s00122-012-1905-8.

Wang, Y.H., H.D. Upadhyaya, A.M. Burrell, S.M. Sahraeian, R.R. Klein and P.E. Klein. 2013. Genetic structure and linkage disequilibrium in a diverse, representative collection of the C4 model plant, Sorghum bicolor. G3 3: 783-793. doi:10.1534/g3.112.004861.

Wendorf, F., A.E. Close, R. Schild, K. Wasylikowa, R.A. Housley, J.R. Harlan and H. Królik. 1992. Saharan exploitation of plants 8,000 years BP. Nature 359: 721-724. doi:10.1038/359721a0.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.L. Jannink, M.E. Sorrells, B. Raman, J.E. Cairns, A. Tarekegne, K. Semagn, Y. Beyene, P. Grudloyma, F. Technow, C. Riedelsheimer and A.E. Melchinger. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2: 1427-1436. doi:10.1534/g3.112.003699.

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, T.T. Tesso, P.S. Schnable, R. Bernardo and J. Yu. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2: 16150. doi:10.1038/nplants.2016.150.

Chapter 3

Multi-trait Regressor Stacking Increased Genomic Prediction Accuracy of Sorghum Grain Composition

Submitted for publication: PLoS One

Citation: Sapkota, S., Boatwright, J. L., Jordan, K. E., Boyles, R., and Kresovich, S. (2020). Multi-trait regressor stacking increased genomic prediction accuracy of sorghum grain composition. *bioRxiv*. 10.1101/2020.04.03.023531

Supplementary file: Appendix B

Multi-trait regressor stacking increased genomic prediction accuracy of sorghum grain composition

Sirjan Sapkota[1,2,*], Jon Lucas Boatwright[1], Kathleen Jordan[1], Richard Boyles[2,3], Stephen Kresovich[1,2],

[1]Advanced Plant Technology Program, Clemson University, Clemson, SC, USA

[2]Department of Plant and Environmental Sciences, Clemson University, Clemson, SC, USA

[3]Pee Dee Research and Education Center, Clemson University, Florence, SC, USA

*correspondence: ssapkot@g.clemson.edu

Abstract

Cereal grains, primarily composed of starch, protein, and fat, are major source of staple for human and animal nutrition. Sorghum, a cereal crop, serves as a dietary staple for over half a billion people in the semi-arid tropics of Africa and South Asia. Genomic prediction has enabled plant breeders to estimate breeding values of unobserved genotypes and environments. Therefore, the use of genomic prediction will be extremely valuable for compositional traits for which phenotyping is labor-intensive and destructive for most accurate results. We studied the potential of Bayesian multi-output regressor stacking (BMORS) model in improving prediction performance over single trait single environment (STSE) models using a grain sorghum diversity panel (GSDP) and a biparental recombinant inbred lines (RILs) population. A total of five highly correlated grain composition traits: amylose, fat, gross energy, protein and starch, with genomic heritability ranging from 0.24 to 0.59 in the GSDP and 0.69 to 0.83 in the RILs were studied. Average prediction accuracies from the STSE model were within a range of 0.4 to 0.6 for all traits across both populations except amylose (0.25) in the GSDP. Prediction accuracy for BMORS increased by 41% and 32% on average over STSE in the GSDP and RILs, respectively. Predicting whole environments by training with remaining environments in BMORS yielded higher average prediction accuracy than from STSE model. Our results show regression stacking methods such as BMORS have potential to accurately predict unobserved individuals and environments, and implementation of such models can accelerate genetic gain.

Introduction

Cereal grains provide more than half of the total human caloric consumption globally and amount to over 80% in some of the poorest nations of the world (Awika, 2011). Sorghum [*Sorghum bicolor* (L.) Moench], a drought-tolerant cereal crop, is a dietary staple for over half a billion people of semi-arid tropics which is inhabited by some of the most food insecure and malnourished populations (Mace et al., 2013). In industrialized countries, such as United States and Australia, grain sorghum is primarily grown for animal feed. But in recent years the uses of sorghum grain have expanded to baking, malting, brewing, and biofortification (Taylor et al., 2006; Taylor, 2012; Zhu, 2014). Therefore, genetic improvement of sorghum grain composition is crucial to mitigate the global malnutrition crisis, to increase efficiency of feed grains used in animal production, and to serve evolving niche markets for gluten-free grains.

In the last two decades, the use of genome-wide markers in prediction of genetic merit of individuals has revolutionized plant and animal breeding. Genomic prediction (GP) uses statistical models to estimate marker effects in a training population with phenotypic and genotypic data which is then used to predict breeding values of individuals solely from genetic markers (Bernardo et al., 2007; Meuwissen et al., 2001). Training population size, genetic relatedness between individuals in training and testing population, marker density, span of linkage disequilibrium and genetic architecture of traits are some of the factors that can affect the predictive ability of the models (Combs et al., 2013; Habier et al., 2007; Zhong et al., 2009). Genomic prediction models are routinely studied and applied by breeding programs around the world in several crops. Novel statistical methods that are capable of incorporating pedigree, genomic, and environmental covariates into statistical-genetic prediction models have emerged as a result of extensive computational research (Crossa et al., 2017).

One of the main advantages of GP is that breeders can use phenotypic values from some lines in some environments to make predictions of new lines and environments. Genomic best linear unbiased prediction (GBLUP) proposed by VanRaden (2008) is probably the most widely used genomic prediction model in both plant and animal breeding. Since then GBLUP model has been extended to include G $\times$ E interactions resulting in improved prediction accuracy of unobserved lines in environments (Burgueño et al., 2012; Jarquın et al., 2014). Burgueño et al. (2012) found an increase in prediction ability of unobserved wheat genotypes by about 20% in multi-environment GBLUP model compared to single environment model. Also an extension of the GBLUP model, Jarquın et al. (2014) introduced a reaction norm model which introduces the main and interaction

effects of markers and environmental covariates by using high-dimensional random variance-covariance structures of markers and environmental covariates. While most of the genomic prediction studies have been on individual traits, breeding programs use selection indices based on several traits to make breeding decisions. To address those challenges, expanded genomic prediction models that perform joint analysis of multiple traits have been studied using empirical and simulated data (Guo et al., 2014a; Jia et al., 2012). Subsequent improvement in prediction accuracy from multi-trait model over single-trait model depends on trait heritability and correlation between the traits involved (Jia et al., 2012; Lado et al., 2018).

Data generated in breeding programs span multiple environment and are recorded for multiple traits for each individual. While multi-environment models and multi-trait models are implemented separately, a single model to account for complexity of variance-covariance structure in a combined multi-trait multi-environment (MTME) model was lacking until Montesinos-López et al. (2016) developed a Bayesian whole genome prediction model to incorporate and analyze multiple traits and multiple environments simultaneously. Montesinos-López et al. (2016) also developed a computationally efficient Markov Chain Monte Carlo (MCMC) method that produces a full conditional distribution of the parameters leading to an exact Gibbs sampling for the posterior distribution. Another MTME model that employs a completely different method was proposed by Montesinos-López et al. (2019). This method, called the Bayesian multi-output regression stacking (BMORS), is a Bayesian version of multi-target regressor stacking (MTRS) originally proposed by Spyromitros-Xioufis et al. (2012, 2016). This method consists of training in two stages: first training multiple learning algorithms for the same dataset and then subsequently combining the predictions to obtain the final predictions.

Genomic prediction for grain quality traits has previously been reported in crops such as wheat (Battenfield et al., 2016; Haile et al., 2018; Heffner et al., 2011), rye (Schulthess et al., 2016), maize (Guo et al., 2014b), and soybean (Duhnen et al., 2017). Hayes et al. (2017) and Battenfield et al. (2016) used near-infrared derived phenotypes in genomic prediction of protein content and end-use quality in wheat. Multi-trait genomic prediction models can simultaneously improve grain yield and protein content despite being negatively correlated (Haile et al., 2018; Rapp et al., 2018). In sorghum, grain macronutrients have shown to be inter-correlated among one another (Boyles et al., 2017), which suggests the multi-trait models may increase predictive ability of individual grain quality traits. The ability to assess genetic merit of unobserved selection candidates across environments is promising for reducing evaluation cost and generation interval in the sorghum breeding pipeline where parental lines of commercial hybrids are currently selected on the basis of extensive progeny testing

(Velazco et al., 2019). In order to extend capacities to performance index selection for multiple environments, we need to study and effectively implement MTME genomic prediction models in our breeding programs. In this study, we report the first implementation of genomic prediction for grain composition in sorghum, and the objective was to assess potential for improvement in prediction accuracy of multi-trait regressor stacking model over single trait model for five grain composition traits: amylose, fat, gross energy, protein and starch.

Materials and methods

Plant material

Grain sorghum diversity panel:

A grain sorghum diversity panel (GSDP) of 389 diverse sorghum accessions was planted in randomized complete block design with two replications in 2013, 2014, and 2017 field seasons at the Clemson University Pee Dee Research and Education Center in Florence, SC. The GSDP included a total of 332 accessions from the original United States sorghum association panel (SAP) developed by Casa et al. (2008). The details on experimental field design and agronomic practices are described in Boyles et al. (2016) and Sapkota et al. (2020). Briefly, the experiments were planted in a two row plots each 6.1 m long, separated by row spacing of 0.762 m with an approximate planting density of 130,000 plants ha^{-1}. Fields were irrigated only when signs of drought stress was seen across the field.

Recombinant inbred population:

A biparental population of 191 recombinant inbred lines (RILs) segregating for grain quality traits was studied along with the GSDP. The parents of the RIL population were BTx642, a yellow-pericarp drought tolerant line, and BTxARG-1, a white pericarp waxy endosperm (low amylose) line. The population was planted in two replicated plots in randomized complete block design across two years (2014 and 2015) in Blackville, SC and College Station, TX. Field design and agronomic practices have previously been described in detail in Boyles et al. (2017).

Phenotyping

The primary panicle of three plants selected from each plot were harvested at physiological maturity. The plants from beginning and end of the row were excluded to account for border effect. Panicles

were air dried to a constant moisture (10-12%) and threshed. A 25g subsample of cleaned and homogenized grain ground to 1-mm particle size with a CT 193 Cyclotec Sample Mill (FOSS North America) was used in near-infrared spectroscopy (NIRS) for compositional analysis.

Grain composition traits such as total fat, gross energy, crude protein, and starch content can be measured using NIRS. Previous studies have shown high NIRS predictability of the traits used in feed analysis (Alencar Figueiredo et al., 2010; Kays et al., 2002). We used a DA 7250^{TM} NIR analyzer (Perten Instruments). The ground sample was packed in a gradually rotating Teflon dish positioned under the instrument's light source and predicted phenotypic values was generated based on calibration curve for spectral measurements. The calibration curve was built using wet chemistry values from a subset of samples. The wet chemistry was performed by Dairyland Laboratories, Inc. (Arcadia, WI) and the Quality Assurance Laboratory at Murphy-Brown, LLC (Warsaw, NC). The details on the prediction curves and wet chemistry can be found in Boyles et al. (2017).

Genotypic data

Genotyping-by-sequencing (GBS) was used for genetic characterization of the GSDP and RILs populations (Boyles et al., 2016, 2017; Morris et al., 2013). Sequenced reads were aligned to the BTx623 v3.1 reference assembly (phytozome) using Burrows-Wheeler aligner (Li et al., 2010). SNP calling, imputation and filtering was done using TASSEL 5.0 pipeline (Glaubitz et al., 2014). The TASSEL plugin FILLIN for GSDP and FSFHap for RILs population were used to impute for missing genotypes. Following imputation SNPs with minor allele frequency (MAF)$<$0.01, and sites missing in more than 10% and 30% of the genotypes in GSDP and RILs, respectively, were filtered. The number of genotypes studied for each population represent those with at least 70% of SNP sites. The genotype matrix with 224,007 SNPs from GSDP and 56,142 SNPs from RILs population was used for whole genome regression.

Statistical analysis

The statistical software environment 'R' was used for model building and analysis (R Core Team, 2019). The phenotypic values of the traits were adjusted for random effects of replications within environment using 'lme4' package in R (Bates et al., 2015). Principal component analysis was done using the R package 'factoextra' (Kassambara et al., 2017). Marker-based estimates of narrow sense (genomic) heritabilities were calculated using the SNP genotype matrix and phenotypic values

using the R package 'heritability' (Kruijer et al., 2015). A matrix with dummy variables '1' and '0' representing combinations of environmental variables (replication and year for GSDP, and replication, year and location for RILs) was used as co-variate in heritability estimation.

Single-trait single-environment (STSE) model:

The following genomic best linear unbiased prediction (GBLUP) model was used to assess prediction performance of an individual trait from a single environment:

$$y_j = \mu + g_j + e_j \tag{1}$$

where y_j is a vector of adjusted phenotypic mean of the jth line (j = 1, 2,..., J). μ is the overall mean which is assigned a flat prior, g_j is a vector of random genomic effect of the jth line, with $\boldsymbol{g} = (g_1, ..., g_j)^T \sim N(\mathbf{0}, \boldsymbol{G}\sigma_1^2)$, σ_1^2 is a genomic variance, $\boldsymbol{G}$ is the genomic relationship matrix in the order $J \times J$ and is calculated (VanRaden, 2008) as $\boldsymbol{G} = \frac{ZZ^T}{2\sum p_j q_j}$, where q_j and p_j denote major and minor allele frequency of jth line respectively, and Z is the design matrix for markers of order $J \times p$ (p is total number of markers). Further, e_j is residual error assigned the normal distribution $e \sim N(0, I\sigma_e^2)$ where I is identity matrix and σ_e^2 is the residual variance with a scaled-inverse Chi-square density.

Bayesian multi-environment (BME) GBLUP model:

Considering genotype × environment interaction can contribute to substantial amount of phenotypic variance in complex traits, we fit the following univariate linear mixed model to account for environmental effects in prediction performance:

$$y_{ij} = E_i + g_j + gE_{ij} + e_{ij} \tag{2}$$

where y_j is a vector of adjusted phenotypic mean of the jth line in the ith environment (i = 1, 2,.., I, j = 1, 2,..., J). E_i represents the effect of ith environment and g_j represents the genomic effect of the jth line as described in equation 1. The term gE_{ij} represents random interaction between the genomic effect of jth line and the ith environment with $gE = (gE_{11}, ..., gE_{IJ})^T \sim N(\mathbf{O}, \sigma_2^2\ I_I \otimes \mathbf{G})$, where σ_2^2 is an interaction variance, and e_{ij} is a random residual associated with the jth line in the ith environment distributed as N(0, σ_e^2) where σ_e^2 is the residual variance.

Bayesian multi-output regressor stacking (BMORS):

BMORS is the Bayesian version of multi-trait (or multi-target) regressor stacking method (Montesinos-López et al., 2019). The multi-target regressor stacking (MTRS) was proposed by Spyromitros-Xioufis et al. (2012, 2016) based on multi-labeled classification approach of Godbole et al. (2004). In BMORS or MTRS, the training is done in two stages. First, L univariate models are implemented using the multi-environment GBLUP model given in equation 2, then instead of using these models for prediction, MTRS performs the second stage of training using a second set of L meta-models for each of the L traits. The following model is used to implement each meta-model:

$$y_{ij} = \beta_1 \hat{Z}_{1ij} + \beta_2 \hat{Z}_{2ij} + ... + \beta_L \hat{Z}_{Lij} + e_{ij} \tag{3}$$

where the covariates $\hat{Z}_{1ij}$, $\hat{Z}_{2ij}$,..., $\hat{Z}_{Lij}$ represent the scaled prediction from the first stage training with the GBLUP model for L traits, and β_1, ..., β_L are the regression coefficients for each covariate in the model. The scaling of each prediction was performed by subtracting its mean (μ_{lij}) and dividing by its corresponding standard deviation (σ_{lij}), that is, $\hat{Z}_{lij}$= ($\hat{y}_{lij}$ - μ_{lij})σ_{lij}^{-1}, for each l = 1,..., L. The scaled predictions of its response variables yielded by the first-stage models as predictor information by the BMORS model. Simply put, the multi-trait regression stacking model is based on the idea that a second stage model is able to correct the predictions of a first-stage model using information about the predictions of other first-stage models (Spyromitros-Xioufis et al., 2012, 2016).

Performance of prediction model:

All prediction models were fit using Bayesian approach in statistical program 'R'. The STSE model (1) was fit using the R package 'BGLR' (Pérez et al., 2014), BME model (2) and BMORS model (3) were fit using the R package 'BMTME' (Montesinos-López et al., 2019). A minimum of 20,000 iterations with 10,000 burn-in steps was used for each Bayesian run.

The evaluation of prediction performance of models was done using a five-fold cross validation (CV), which means 80% of the samples were used as training set and testing was done on the remaining 20% for each cross-validation fold. The individuals were randomly assigned into five mutually exclusive folds. Four folds were used to train prediction models and to predict the genomic estimated breeding values (GEBVs) of the individuals in fifth fold (validation/test set). The accuracy of prediction for each fold was calculated as Pearson's correlation coefficient (r) between predicted values and adjusted phenotypic means for the individuals in validation set. Each cross validation run,

therefore, resulted in five estimates of prediction accuracy. The same set of individuals were assigned to training and validation across different traits and models tested by using *set.seed()* function in R. In order to avoid bias due to sampling, we performed 10 different cross-validation runs to calculate the mean and dispersion of the prediction accuracies.

Results

Phenotypic variation

A single calibration curve for NIRS was used for both populations studied. Table 1 outlines the summary statistics of NIRS predictions and phenotypic distribution and heritability of the grain composition traits. The cross validation accuracy (R^2) of the NIRS calibration curve was moderately high to high, except for fat which had a moderate R^2 value (0.41). We had a total of three environments (three years in one location) for the GSDP and four environments (two years in two locations) for the RILs. Traits were normally distributed except amylose in two 2014 environments in the RILs which had bimodal distribution (S1 Fig, S2 Fig). All traits showed significant variation in distribution across the environments, except starch in GSDP.

Table 1. Summary statistics of near infrared spectroscopy (NIRS) calibration and phenotypic distribution in grain sorghum diversity panel (GSDP) and recombinant inbred lines (RILs). R^2 is the prediction accuracy and SECV is the standard error of cross validation for the NIRS calibration curve. Mean represents the phenotypic mean of the trait with its standard deviation (SD). h^2 is the estimate of genomic heritability.

Trait	NIRS		GSDP		RILs	
	R^2	SECV	Mean ± SD	h^2	Mean ± SD	h^2
Amylose	0.60	2.24	13.87 ± 2.98	0.24	11.49 ± 4.32	0.77
Fat	0.41	0.53	2.53 ± 0.57	0.54	3.07 ± 0.67	0.76
Gross energy	0.71	25.80	4108.33 ± 55.15	0.59	4124.56 ± 41.74	0.69
Protein	0.96	0.27	12.02 ± 1.45	0.39	11.43 ± 1.03	0.83
Starch	0.89	0.75	68.30 ± 2.44	0.44	68.37 ± 1.87	0.79

The genomic heritabilities of all traits except gross energy were significantly higher ($p < 0.05$) in the RILs than in the GSDP (Table 1). Trait heritabilities were high in the RILs, with protein and gross energy having the highest and lowest heritabilities, respectively. In the GSDP, genomic heritability was moderately high for fat and gross energy, moderate for protein and starch, and low for amylose. The poor genomic heritability (0.24) of amylose in the GSDP was expected because only a very small proportion (1%) of accessions have low amylose as a result of *waxy* gene (Mendelian trait).

Fig 1 shows correlation between the adjusted phenotypic means for trait and environment combination. Starch was negatively correlated ($p<0.001$) with all other traits in both populations except for amylose in the RILs. Fat, protein and gross energy were significantly positively ($p<0.001$) correlated to each other across environments in both populations. The strongest positive correlation was between gross energy and fat, whereas the strongest negative correlations were found between starch $\sim$ gross energy and starch $\sim$ protein. Moderate to high positive correlation was observed between years for all traits (Fig 1). We conducted a principal component analysis (PCA) of correlation matrix for the traits in each environment. In both populations, the first component separated amylose and starch from the other three traits, whereas, the second component separated amylose from starch and gross energy from protein and fat (S3 Fig). The first component explained 78.8% and 75.9% of variation, and second component explained 6.3% and 9.8% of variation in the GSDP and RILs, respectively. The third principal component in the RILs separated proteins from fat and explained about 7.6% of the variation.

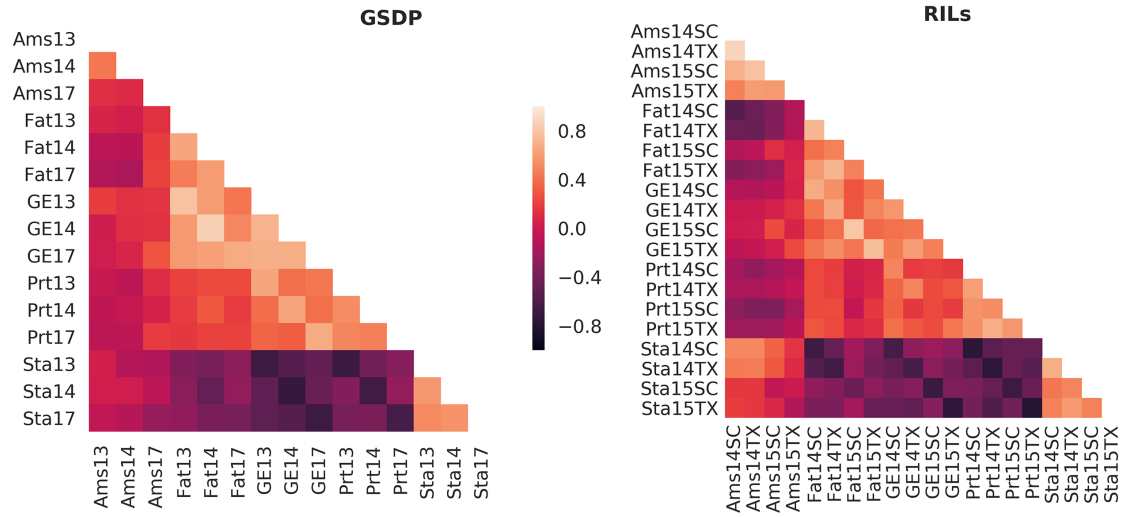


Figure 1. Correlation between traits across year and location combination for the two populations. Ams: amylose, GE: gross energy, Prt: protein, Sta: starch, SC: South Carolina, TX: Texas, and numbers in x and y-axes represent the year.

Prediction performance in single and multiple environment

We first implemented GBLUP prediction model for single-trait single-environment (STSE). Prediction accuracies of the STSE model varied across environments in both populations (Fig 2). The environments 2014 in the GSDP and TX2014 in the RILs had highest average prediction accuracy but were not always the best predicted environment for all traits. While poorly predicted for amylose,

the environments 2017 in the GSDP and TX2015 in the RILs had higher prediction accuracy for starch compared to all or most environments. Despite variation across environments and populations, the average prediction accuracies from the STSE were within the range of 0.4 to 0.6 for all traits except amylose (0.25) in the GSDP. The average prediction accuracy of the STSE model in the GSDP was positively correlated (r=0.86) with the genomic heritability of the traits. In the RILs, there was a positive correlation (r=0.77) between average prediction accuracy and genomic heritability for amylose, fat and gross energy, but the traits (protein and starch) with the highest heritabilities had relatively lower average prediction accuracies.

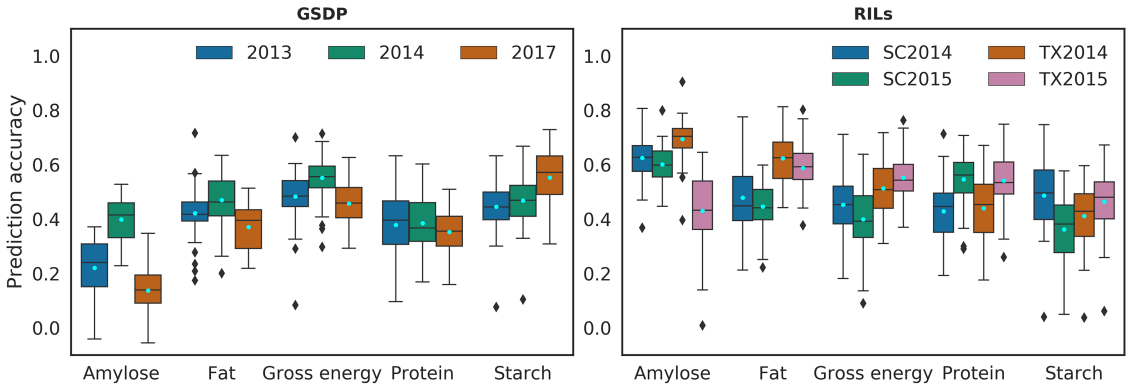


Figure 2. Prediction accuracy for single-trait single-environment model. Legend represents the environment/years. SC: South Carolina, TX: Texas. Pale blue dots represent the mean of prediction accuracy.

We didn't see substantial improvement in multi-environment (BME) model over the STSE prediction accuracies (Fig 3). In the GSDP, the multi-environment models resulted in a decline in average prediction accuracy compared to the STSE model for fat (21%), amylose (10%) and protein (4%), however, no significant change was observed for gross energy and starch (S4 Fig, S1 Table). The prediction accuracy in the RILs increased by an average of 3% in the BME compared to the STSE, however, the overall trend of prediction accuracy for traits and environments remained unchanged (S4 Fig). The environment SC2014 showed consistent increase in accuracy for BME over STSE model across all traits with about 10% increase for protein (S2 Tab). Amylose in TX2015 environment had the single greatest increase (12%) in average prediction accuracy in the BME among all trait-environment combinations for the RILs.

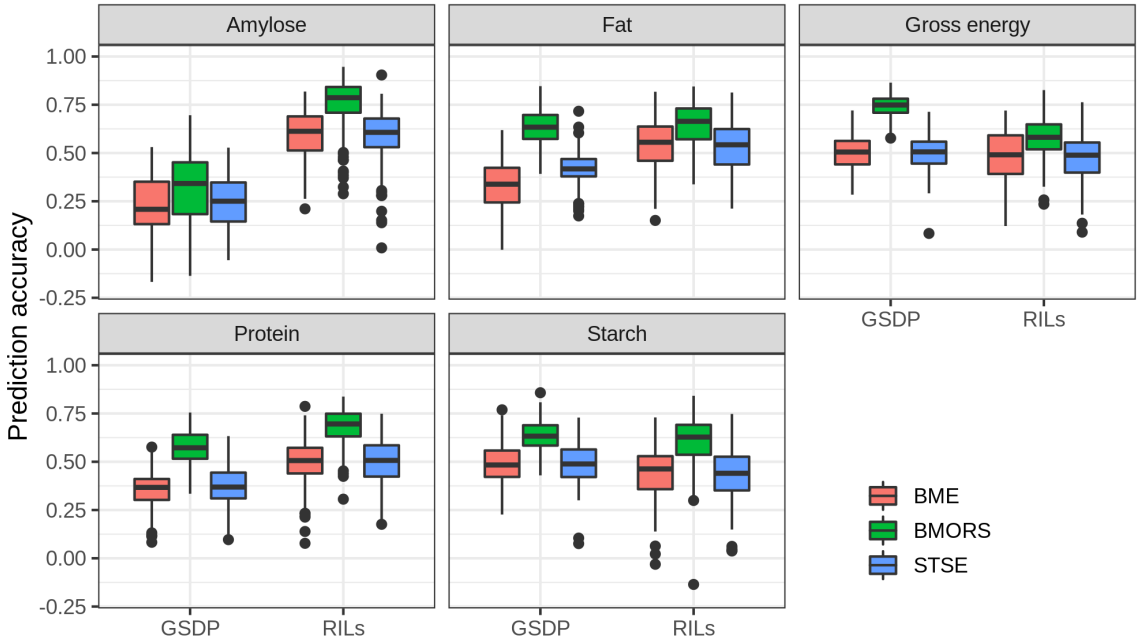


Figure 3. Average prediction accuracy of traits for the three prediction methods in the two populations.

Bayesian multi-output regression stacking

We tested two different prediction schemes in the BMORS prediction model using the two functions *BMORS()* and *BMORS_Env()* as described in Montesinos-López et al. (2019). While the *BMORS()* function was used for a five-fold CV as described in the methods section, the *BMORS_Env()* was used to assess the prediction performance of whole environments while using the remaining environments as training.

Five-fold CV

The prediction accuracy from five-fold CV in BMORS increased by 41% and 32% on average over the STSE model in GSDP and RILs, respectively. Fig 4 shows the prediction accuracy of BMORS for each trait and environment combination across the two populations. While the percent change in accuracy varied across environments, the BMORS model nonetheless had higher average prediction accuracy than the STSE and BME models for all traits (Fig 3). The increase in average accuracy ranged from 11% (amylose, 2014) to 66% (amylose, 2013) in the GSDP with exception of amylose in 2017 (13% decrease), and 15% (fat, SC2015) to 60% (protein, TX2014) in the RILs (S1 Table). The increase in average prediction accuracy was higher (35%) for both locations in 2014 for the RILs,

whereas, the year 2013 in the GSDP increased the most (S1 Table, S2 Table). Among the traits, protein (54%) had the greatest average increase in prediction accuracy in the GSDP, whereas in the RILs, protein and starch (42%) both showed the greatest increase.

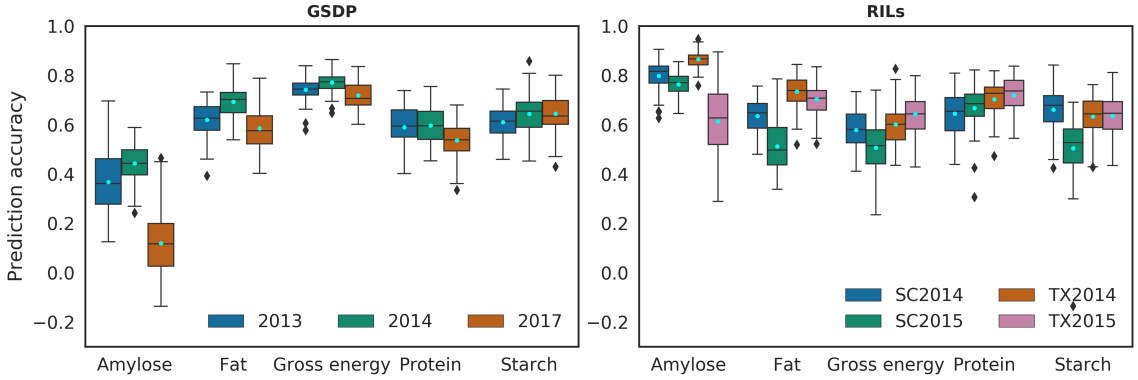


Figure 4. Prediction accuracy of Bayesian multi-output regressor stacking (BMORS) model using five-fold CV. Legend represents the years/environment. SC: South Carolina, TX: Texas. Pale blue dots show mean of the prediction accuracy.

Prediction of whole environment

Predicting a whole environment using the BMORS model usually yielded higher accuracy than the mean prediction accuracy from the STSE model for each trait and environment combination (Fig 2, 5, Table 2). The distribution of prediction accuracy across trait and environment combination were, however, similar to the results from the STSE model. In the GSDP, little variation in prediction accuracies was observed across environments for gross energy, starch and protein, whereas, amylose and fat showed greater variability in prediction accuracy between environments. In the RILs, prediction accuracy for all traits except protein had high variability across the environments (Table 2).

In order to assess predictability by location or year in the RILs, we tested one location or year by training the BMORS model using the other location or year, respectively (Table 2). The Texas location had higher accuracy of prediction for fat (+0.11) and gross energy (+0.1) compared to South Carolina, but rest of the traits had similar prediction accuracy (difference <0.02). Prediction accuracy of whole years varied across traits, amylose (+0.09) and fat(+0.04) were higher in 2014, protein was higher (+0.05) in 2015, and starch and gross energy were similar.

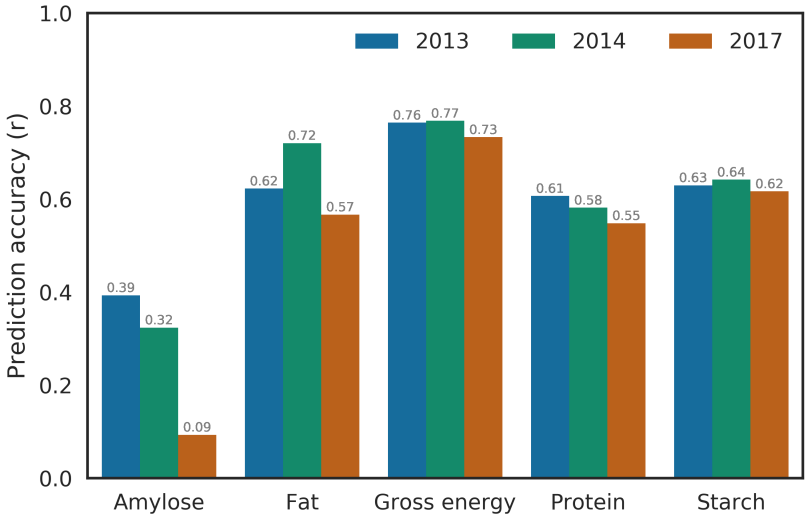


Figure 5. Prediction accuracy of the test environments predicted using the Bayesian multi-output regressor stacking (BMORS) for whole environment in the diversity panel. Values on top of the bar represent the height of the bar.

Table 2. Prediction accuracy of the test environments predicted using the BMORS_Env in the RILs.

Trait	Year × Location				Location		Year	
	SC2014	SC2015	TX2014	TX2015	SC	TX	2014	2015
Amylose	0.79	0.80	0.88	0.60	0.76	0.74	0.74	0.65
Fat	0.69	0.49	0.78	0.74	0.60	0.71	0.64	0.60
Gross energy	0.56	0.49	0.62	0.66	0.48	0.58	0.56	0.56
Protein	0.65	0.66	0.66	0.70	0.59	0.58	0.61	0.66
Starch	0.64	0.52	0.68	0.60	0.55	0.56	0.56	0.55

Discussion

Phenotyping for grain compositional traits is: 1) challenging and labor-intensive, 2) destructive for most accurate results, and 3) only performed after plants reach physiological maturity and are harvested. The use of genomic prediction for compositional traits will be extremely valuable because it increases selection intensity and decreases generational interval by overcoming the phenotyping challenges. Moreover, these traits are complex and quantitatively inherited so will benefit from genomic prediction's ability to account for many small effect QTLs in estimating breeding values.

Trait architecture and prediction accuracy

While the accuracy of NIRS calibration for traits in this study ranged from moderate to high, there was prediction error associated with NIRS prediction. However, it is unclear if and what effects NIRS prediction error had on genomic prediction. No direct correlation was observed between the genomic prediction accuracy and NIRS statistics for the traits studied. The trait with the lowest NIRS R^2, fat, was predicted as well as or better than starch, protein and gross energy, which had NIRS R^2 >0.7. Despite varying strength of correlations between traits across the two populations studied, the nature of relationship was similar for a given pair of traits, which is also in agreement with previous studies (Boyles et al., 2017; Murray et al., 2008; Sukumaran et al., 2012). The strong negative relationship of starch and amylose to protein, fat and gross energy was further elucidated by the PCA analysis of phenotypic correlation matrix (S3 Fig). Since starch, protein and fat were measured on a percent dry matter basis, the strong correlation between them is expected.

Genetic relatedness and trait architecture are known to affect the accuracy of genomic prediction (Habier et al., 2007; Jannink et al., 2010). The genetic relatedness between individuals and heritability of the traits were higher in the RILs than in GSDP (S3 Fig, Table 1). Those factors could be contributing to higher average prediction accuracy in the RILs. However, the average prediction accuracies for gross energy and starch were comparable between GSDP and RILs (Fig 3). Prediction accuracy in the GSDP could have been boosted by greater genetic diversity despite lower genetic relatedness (Sapkota et al., 2020). Heffner et al. (2011) observed a prediction accuracy of 0.5-0.6 for wheat flour protein in two biparental populations. Guo et al. (2014b) reported prediction accuracies of 0.44 and 0.8 for protein and amylose in rice diversity panel. Similar results were observed in our STSE models for protein content (Fig 2). Whereas, lower prediction accuracy of amylose in our diversity panel is probably due to the lack of sufficient low-amylose lines with the *waxy* gene (Boyles et al., 2017). While we couldn't find any previous genomic prediction studies on starch, fat and gross energy, these traits are nutritionally one of the most important traits in cereal grains. The moderate to high prediction accuracy observed suggests implementation of genomic selection can improve genetic gain for these grain quality traits.

Multi-trait regressor stacking

One of the daunting tasks of genomic prediction is estimating the effects of unobserved individuals and environments. As multiple traits are analyzed across several environments, the ability to combine

information from multiple traits and environments can be crucial in increasing accuracy of prediction (Burgueño et al., 2012; Guo et al., 2014a; Jia et al., 2012). When the correlations among traits are high, prediction accuracies of complex traits can be increased by using multivariate model that takes this correlation into account (Jia et al., 2012; Montesinos-López et al., 2016). We fit a Bayesian multi-environment (BME) model (2) that takes the genotype × environment effects into consideration. In the GSDP, where environments were three years at the same location, the BME model showed a slight decline (7%) in average prediction accuracy which was mostly due to the two traits, amylose and fat (Supplementary Table S1). The RILs showed slight increase (2-3%) in prediction accuracy of traits when averaged over the environments, but there was variability across the environments (S2 Table).

We implemented two functions [*BMORS*() and *BMORS_Env*()] which are not only used to evaluate prediction accuracy but are also computationally efficient (Montesinos-López et al., 2019). The BMORS model (3) performs two-stage training by stacking the multi-environment models from all the traits. The five-fold cross validation conducted for BMORS was similar to the CV1 strategy of Montesinos-López et al. (2016). The use of multi-trait models has been consistently shown to increase prediction accuracy over single-trait models across different crops and traits (Bhatta et al., 2020; Guo et al., 2014a; Jia et al., 2012; Lado et al., 2018). The multi-target regressor stacking increased average prediction accuracy by 41% and 32% in the GSDP and RILs, respectively, as compared to the STSE prediction accuracy. Average prediction accuracy of all traits improved in BMORS over STSE and BME across both the populations (Fig 3). Consistent improvement in accuracy of BMORS is a result of the ability to use not only correlation between traits but also between environment in the model training (Montesinos-López et al., 2016, 2019). The ability to accurately estimate genetic merit of lines in unobserved environments is of tremendous value in plant breeding. Our results show potential of *BMORS_Env()* function for predicting the whole environment. Testing a whole environment by training BMORS model using all other environments resulted in higher prediction accuracy for that trait-environment combination than using STSE or BME model. Prediction accuracy of all environments were 0.5 or higher with exception of amylose in GSDP, the reason for which we have discussed above (Fig 5, Table 2).

Application for breeding

Grain quality traits such as starch and protein content have been under selection since the inception of phenotypic selection in modern breeding practices. More recently, total energy supplement of grain has gained attention for increasing feed efficiency in animal production, and a need exists for increasing total calories for human nutrition in the wake of global malnutrition crisis. Despite high correlations among these traits, the genetic variation underlying starch, protein and fat can be decoupled. Boyles et al. (2017) have shown major and minor effect QTLs underlying the three traits are distributed across the genome and are segregating in biparental populations. However, in practice, selection would be conducted simultaneously for these traits using a selection index rather than for individual traits. Velazco et al. (2019) observed an increase in predictive ability by using a multi-trait model for grain yield and stay green in sorghum, and argue that such an exercise would allow for using selection index for implementation of genomic selection for correlated traits. Increased prediction accuracy, improved selection index, and estimation of precise genetic, environmental and residual co-variances makes multi-trait multi-environment models preferable over univariate models (Montesinos-López et al., 2016). The multi-trait regression stacking model we tested shows large scale improvement in model prediction and can be used in tandem with Bayesian multi-trait multi-environment (BMTME) model for parameter estimation and assessing prediction accuracy. The ability to estimate genetic effects and breeding values of unobserved environments will be of great advantage to predict performance in diverse environments and for implementation of selection theory.

Conclusion

Phenotyping of grain compositional traits using near-infrared spectroscopy is labor-intensive, generally destructive, and time limiting. Therefore, the use of genomic selection for these traits will be extremely valuable. This study establishes the potential to improve genomics-assisted selection of grain composition traits by using multi-trait multi-environment model. The phenotypic measurements obtained from NIRS prediction were amenable to genomic selection as shown by moderate to high prediction accuracy for single trait prediction. While multi-environment model alone didn't lead to much improvement over single environment model, stacking of regression from multiple traits showed substantial improvement in prediction accuracy. The prediction accuracy increased by 32% and 41% in the RILs and GSDP, respectively, when using the Bayesian multi-output regressor stacking

(BMORS) model compared to a single trait single environment model. The ability to predict line performance in an unobserved environment is of great importance to breeding programs, and results show high accuracy for predicting whole environments using BMORS.

Supporting information

S1 Fig. Phenotypic distribution of grain composition traits in the RILs. In the x-axes, SC: South Carolina, TX: Texas, numbers represent years. Values are percentage dry basis for protein, fat and starch; gross energy is in KCal/lb; and amylose is in percent of starch.

S2 Fig. Phenotypic distribution of grain composition traits in the GSDP. Numbers in x-axes represent years. Values are percentage dry basis for protein, fat and starch; gross energy is in Cal/g; and amylose is in percent of starch.

S3 Fig. PCA analysis of correlation matrix between traits. a. GSDP, and **b**. RILs. Ams: amylose, GE: gross energy, Prt: protein, Sta: starch, SC: South Carolina, TX: Texas. The numbers in the text represent years of the environment.

S4 Fig. Prediction accuracy using five-fold CV in Bayesian multi-environment (BME) model. a. GSDP, and **b**. RILs. Legend represents the environment/years. SC: South Carolina, TX: Texas. Pale blue dots represent the mean of prediction accuracy.

S5 Fig. Heatmap for genomic relationship matrix calculated using vanRaden (2008). a. GSDP, **b**. RILs. Trees show hierarchical clustering using Euclidean distance.

S1 Table. Percent change in prediction accuracy over the single trait single environment model (STSE) model in the GSDP. BME: Bayesian multi-environment, and BMORS: Bayesian multi-output regressor stacking.

S2 Table. Percent change in prediction accuracy over the single trait single environment model (STSE) model in the RILs. BME: Bayesian multi-environment, and BMORS: Bayesian multi-output regressor stacking.

Acknowledgments

The authors would like to thank William L. Rooney and Brian K. Pfeiffer for their contributions to phenotyping of the recombinant inbred population at College Station, TX. Our appreciation goes to the Wade Stackhouse Fellowship, and Robert and Lois Coker Endowment for their support during the study. Clemson University's computing cluster, Palmetto, was used for intensive data analyses. This study was funded by research grants from US Department of Energy ARPA-E TERRA (Award# DE-AR0000595, and DE-AR0001134).

Data availability

The codes and data used in the study are available at github.com/sirjansapkota/GrainComp_GS.

References

Alencar Figueiredo, Lucio Flavio de, Bassirou Sine, Jacques Chantereau, Christian Mestres, Geneviève Fliedel, J-F Rami, J-C Glaszmann, Monique Deu, and Brigitte Courtois (2010). "Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI, Ae1, Wx and O2". In: *Theoretical and applied genetics* 121.6, pp. 1171–1185.

Awika, Joseph M (2011). "Major cereal grains production and use around the world". In: *Advances in cereal science: implications to food processing and health promotion*. ACS Publications, pp. 1–13.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Battenfield, Sarah D, Carlos Guzmán, R Chris Gaynor, Ravi P Singh, Roberto J Peña, Susanne Dreisigacker, Allan K Fritz, and Jesse A Poland (2016). "Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program". In: *The plant genome* 9.2.

Bernardo, Rex and Jianming Yu (2007). "Prospects for genomewide selection for quantitative traits in maize". In: *Crop Science* 47.3, pp. 1082–1090.

Bhatta, Madhav, Lucia Gutierrez, Lorena Cammarota, Fernanda Cardozo, Silvia Germán, Blanca Gómez-Guerrero, Maria Fernanda Pardo, Valeria Lanaro, Mercedes Sayas, and Ariel J Castro (2020). "Multi-trait Genomic Prediction Model Increased the Predictive Ability for Agronomic

and Malting Quality Traits in Barley (Hordeum vulgare L.)" In: *G3: Genes, Genomes, Genetics* 10.3, pp. 1113–1124.

Boyles, Richard E, Elizabeth A Cooper, Matthew T Myers, Zachary Brenton, Bradley L Rauh, Geoffrey P Morris, and Stephen Kresovich (2016). "Genome-wide association studies of grain yield components in diverse sorghum germplasm". In: *The plant genome* 9.2.

Boyles, Richard E, Brian K Pfeiffer, Elizabeth A Cooper, Bradley L Rauh, Kelsey J Zielinski, Matthew T Myers, Zachary Brenton, William L Rooney, and Stephen Kresovich (2017). "Genetic dissection of sorghum grain quality traits using diverse and segregating populations". In: *Theoretical and applied genetics* 130.4, pp. 697–716.

Burgueño, Juan, Gustavo de los Campos, Kent Weigel, and José Crossa (2012). "Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers". In: *Crop Science* 52.2, pp. 707–719.

Casa, Alexandra M, Gael Pressoir, Patrick J Brown, Sharon E Mitchell, William L Rooney, Mitchell R Tuinstra, Cleve D Franks, and Stephen Kresovich (2008). "Community resources and strategies for association mapping in sorghum". In: *Crop science* 48.1, pp. 30–40.

Combs, Emily and Rex Bernardo (2013). "Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers". In: *The Plant Genome* 6.1.

Crossa, José, Paulino Pérez-Rodrıguez, Jaime Cuevas, Osval Montesinos-López, Diego Jarquın, Gustavo de los Campos, Juan Burgueño, Juan M González-Camacho, Sergio Pérez-Elizalde, Yoseph Beyene, et al. (2017). "Genomic selection in plant breeding: methods, models, and perspectives". In: *Trends in plant science* 22.11, pp. 961–975.

Duhnen, Alexandra, Amandine Gras, Simon Teyssèdre, Michel Romestant, Bruno Claustres, Jean Daydé, and Brigitte Mangin (2017). "Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy". In: *Crop Science* 57.3, pp. 1325–1337.

Glaubitz, Jeffrey C, Terry M Casstevens, Fei Lu, James Harriman, Robert J Elshire, Qi Sun, and Edward S Buckler (2014). "TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline". In: *PloS one* 9.2, e90346.

Godbole, Shantanu and Sunita Sarawagi (2004). "Discriminative methods for multi-labeled classification". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 22–30.

Guo, Gang, Fuping Zhao, Yachun Wang, Yuan Zhang, Lixin Du, and Guosheng Su (2014a). "Comparison of single-trait and multiple-trait genomic prediction models". In: *BMC genetics* 15.1, p. 30.

Guo, Zhigang, Dominic M Tucker, Christopher J Basten, Harish Gandhi, Elhan Ersoz, Baohong Guo, Zhanyou Xu, Daolong Wang, and Gilles Gay (2014b). "The impact of population structure on genomic prediction in stratified populations". In: *Theoretical and applied genetics* 127.3, pp. 749–762.

Habier, David, Rohan L Fernando, and Jack CM Dekkers (2007). "The impact of genetic relationship information on genome-assisted breeding values". In: *Genetics* 177.4, pp. 2389–2397.

Haile, Jemanesh K, Amidou N'Diaye, Fran Clarke, John Clarke, Ron Knox, Jessica Rutkoski, Filippo M Bassi, and Curtis J Pozniak (2018). "Genomic selection for grain yield and quality traits in durum wheat". In: *Molecular breeding* 38.6, p. 75.

Hayes, BJ, J Panozzo, CK Walker, AL Choy, S Kant, D Wong, J Tibbits, HD Daetwyler, S Rochfort, MJ Hayden, et al. (2017). "Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes". In: *Theoretical and applied genetics* 130.12, pp. 2505–2519.

Heffner, Elliot L, Jean-Luc Jannink, Hiroyoshi Iwata, Edward Souza, and Mark E Sorrells (2011). "Genomic selection accuracy for grain quality traits in biparental wheat populations". In: *Crop Science* 51.6, pp. 2597–2606.

Jannink, Jean-Luc, Aaron J Lorenz, and Hiroyoshi Iwata (2010). "Genomic selection in plant breeding: from theory to practice". In: *Briefings in functional genomics* 9.2, pp. 166–177.

Jarquın, Diego, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. (2014). "A reaction norm model for genomic selection using high-dimensional genomic and environmental data". In: *Theoretical and applied genetics* 127.3, pp. 595–607.

Jia, Yi and Jean-Luc Jannink (2012). "Multiple-trait genomic selection methods increase genetic value prediction accuracy". In: *Genetics* 192.4, pp. 1513–1522.

Kassambara, Alboukadel and Fabian Mundt (2017). "Factoextra: extract and visualize the results of multivariate data analyses". In: *R package version* 1.4.

Kays, Sandra E and Franklin E Barton (2002). "Rapid prediction of gross energy and utilizable energy in cereal food products using near-infrared reflectance spectroscopy". In: *Journal of agricultural and food chemistry* 50.5, pp. 1284–1289.

Kruijer, Willem, Martin P Boer, Marcos Malosetti, Pádraic J Flood, Bas Engel, Rik Kooke, Joost JB Keurentjes, and Fred A van Eeuwijk (2015). "Marker-based estimation of heritability in immortal populations". In: *Genetics* 199.2, pp. 379–398.

Lado, Bettina, Daniel Vázquez, Martin Quincke, Paula Silva, Ignacio Aguilar, and Lucia Gutiérrez (2018). "Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality". In: *Theoretical and Applied Genetics* 131.12, pp. 2719–2731.

Li, Heng and Richard Durbin (2010). "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5, pp. 589–595.

Mace, Emma S, Shuaishuai Tai, Edward K Gilding, Yanhong Li, Peter J Prentis, Lianle Bian, Bradley C Campbell, Wushu Hu, David J Innes, Xuelian Han, et al. (2013). "Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum". In: *Nature communications* 4, p. 2320.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps". In: *Genetics* 157.4, pp. 1819–1829. ISSN: 0016-6731. URL: `https://www.genetics.org/content/157/4/1819`.

Montesinos-López, Osval A, Abelardo Montesinos-López, José Crossa, Fernando H Toledo, Oscar Pérez-Hernández, Kent M Eskridge, and Jessica Rutkoski (2016). "A genomic Bayesian multi-trait and multi-environment model". In: *G3: Genes, Genomes, Genetics* 6.9, pp. 2725–2744.

Montesinos-López, Osval A, Abelardo Montesinos-López, Francisco Javier Luna-Vázquez, Fernando H Toledo, Paulino Pérez-Rodrıguez, Morten Lillemo, and José Crossa (2019). "An R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction". In: *G3: Genes, Genomes, Genetics* 9.5, pp. 1355–1369.

Morris, Geoffrey P, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. (2013). "Population genomic and genome-wide association studies of agroclimatic traits in sorghum". In: *Proceedings of the National Academy of Sciences* 110.2, pp. 453–458.

Murray, Seth C, Arun Sharma, William L Rooney, Patricia E Klein, John E Mullet, Sharon E Mitchell, and Stephen Kresovich (2008). "Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates". In: *Crop Science* 48.6, pp. 2165–2179.

Pérez, Paulino and Gustavo de Los Campos (2014). "Genome-wide regression and prediction with the BGLR statistical package". In: *Genetics* 198.2, pp. 483–495.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rapp, M, V Lein, F Lacoudre, J Lafferty, E Müller, G Vida, V Bozhanova, A Ibraliu, P Thorwarth, HP Piepho, et al. (2018). "Simultaneous improvement of grain yield and protein content in durum wheat by different phenotypic indices and genomic selection". In: *Theoretical and applied genetics* 131.6, pp. 1315–1329.

Sapkota, Sirjan, Rick Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich (2020). "Impact of sorghum racial structure and diversity on genomic prediction of grain yield components". In: *Crop Science* 60.1, pp. 132–148. DOI: 10.1002/csc2.20060.

Schulthess, Albert Wilhelm, Yu Wang, Thomas Miedaner, Peer Wilde, Jochen C Reif, and Yusheng Zhao (2016). "Multiple-trait and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes". In: *Theoretical and Applied Genetics* 129.2, pp. 273–287.

Spyromitros-Xioufis, Eleftherios, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas (2012). "Multi-label classification methods for multi-target regression". In: *arXiv preprint arXiv:1211.6581*, pp. 1159–1168.

Spyromitros-Xioufis, Eleftherios, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas (2016). "Multi-target regression via input space expansion: treating targets as inputs". In: *Machine Learning* 104.1, pp. 55–98.

Sukumaran, Sivakumar, Wenwen Xiang, Scott R Bean, Jeffrey F Pedersen, Stephen Kresovich, Mitchell R Tuinstra, Tesfaye T Tesso, Martha T Hamblin, and Jianming Yu (2012). "Association mapping for grain quality in a diverse sorghum collection". In: *The Plant Genome* 5.3, pp. 126–135.

Taylor, John RN, Tilman J Schober, and Scott R Bean (2006). "Novel food and non-food uses for sorghum and millets". In: *Journal of cereal science* 44.3, pp. 252–271.

Taylor, JRN (2012). "Food product development using sorghum and millets: opportunities and challenges". In: *Quality Assurance and Safety of Crops & Foods* 4.3, pp. 151–151.

VanRaden, Paul M (2008). "Efficient methods to compute genomic predictions". In: *Journal of dairy science* 91.11, pp. 4414–4423.

Velazco, Julio G, David R Jordan, Emma S Mace, Colleen H Hunt, Marcos Malosetti, and Fred A Van Eeuwijk (2019). "Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis". In: *Frontiers in plant science* 10.

Zhong, Shengqiang, Jack CM Dekkers, Rohan L Fernando, and Jean-Luc Jannink (2009). "Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study". In: *Genetics* 182.1, pp. 355–364.

Zhu, Fan (2014). "Structure, physicochemical properties, modifications, and uses of sorghum starch". In: *Comprehensive Reviews in Food Science and Food Safety* 13.4, pp. 597–610.

Chapter 4

Genome-wide Association and Gene Network Analysis for Starch and Protein in Sorghum

Manuscript in preparation for submission.

Supplementary file: Appendix C

Genome-wide association and gene network analysis for starch and protein in sorghum

Sirjan Sapkota[1,2,*], Jon Lucas Boatwright[2], Kathleen Jordan[2], Richard Boyles[1,3], Stephen Kresovich[1,2],

1 Department of Plant and Environmental Sciences, Clemson University, Clemson, SC, USA

2 Advanced Plant Technology Program, Clemson University, Clemson, SC, USA

3 Pee Dee Research and Education Center, Clemson University, Florence, SC, USA

* correspondence: ssapkot@g.clemson.edu

Abstract

The grains of cereals are ultimate sink for macromolecules such as starch and protein which serve as principal source of nutrition. Dissection of genetic basis of phenotypic variation for starch and protein is important in metabolic engineering of grain. Genetic studies of starch and protein in sorghum, a cereal crop, has involved single traits and metabolic network involved in their regulation are not completely characterized. In this study we used univariate and multivariate (MV) linear mixed models (LMM) to identify associated genomic regions, potential candidate genes and their interactors. Six single nucleotide polymorphism (SNPs) in strong linkage disequilibrium (r^2 >0.8) from ∼52 Mb of chromosome 8 were significantly associated with starch content. Five of those SNPs were located within mRNA of a heat shock protein 90 (HSP90), *Sobic.008G111600*, with two of them in the coding sequences of the gene. The HSP90 had a total of 142 high confidence (PPI-score: 0.6) first interactors and the network was enriched for six biochemical pathways including protein processing and export. A SNP, S4_60623675, identified using MV-LMM model was located at 5'UTR of a fatty acid desaturase gene, *Sobic.004G260800*, which interacted with another fatty acid desaturase and several nitrate reductase genes. The two candidates, HSP90 and FAD2, were found to be highly expressed in reproductive tissues. We conclude multivariate analysis of correlated phenotypes can identify biologically important metabolic networks and functional analyses of identified gene candidates can be beneficial in understanding grain filling in sorghum and other cereals.

Introduction

The seeds of cereals, that represent an important sink for metabolites during grain filling, are principal source of human and animal nutrition. Sorghum [*Sorghum bicolor*. (L.) Moench] is a cereal crop that provides dietary staple for over half a billion people in semi-arid tropics (Mace et al., 2013). While primarily used as animal feed in industrialized economies, the end use products of sorghum grain has diversified to include baking, malting, brewing, and bio-fortification (Zhu, 2014). Understanding the genetic basis of phenotypic variation in grain composition such as starch and protein could provide the basis for metabolic engineering of these macromolecules through selective breeding.

Linkage mapping has been a powerful method to identify quantitative trait loci (QTL) that cosegregate with a given trait but suffers from two fundamental limitations; only allelic diversity that segregates between the parents can be assayed, and the amount of recombination from bi-parental crosses places a limit on the mapping resolution (Korte et al., 2013). In contrast, genome-wide association studies (GWAS) have mapped genetic variants associated to phenotypes to a much higher resolution using whole genome markers in a diverse group of individuals. The cost effectiveness in generating large scale genotypic data has now led to swathe of GWAS in crops and focus has shifted towards computational challenges (Myles et al., 2009). Most application of GWAS has focused on single traits, whereas phenotypes are usually correlated and might be controlled by genetic loci with pleiotropic effects. Meanwhile, studies have shown that joint analysis of correlated phenotypes can exploit the correlation among the phenotypes for detecting additional genetic variants with small effects across multiple traits (Korte et al., 2012; Thoen et al., 2017; Carlson et al., 2019; Rice et al., 2020). Some of the approaches to leverage correlation between traits in association analysis include: use of ratios of directly related traits in univariate GWAS (Gieger et al., 2008), combining test statistics from univariate GWAS of each trait to detect pleiotropic effects (Yang et al., 2010), using dimension reduction technique to derive transformed phenotypes for univariate GWAS (Aschard et al., 2014), and directly modeling multiple traits into a multivariate linear mixed models (Korte et al., 2012; Zhou et al., 2014).

Association studies for grain quality has been reported in several cereal crops such as maize (Wilson et al., 2004; Cook et al., 2012), rice (Zhao et al., 2011; Wang et al., 2017), sorghum (Sukumaran et al., 2012; Rhodes et al., 2017), and wheat (Reif et al., 2011; Gaire et al., 2019). In diverse sorghum accessions, starch and protein show continuous variation ranging from 60 to 72% and 8 to 18% of total grain, respectively (Rhodes et al., 2017). Previous association analyses for starch and protein

content have identified significantly associated genomic regions in sorghum (Sukumaran et al., 2012; Rhodes et al., 2017; Boyles et al., 2017). Starch and protein content represent majority of grain composition and are likely to be controlled by genetic loci with pleiotropic effects. Such genetic loci could have gone undetected in single trait GWAS, and multivariate GWAS might be able to identify such associations. Furthermore, the path from genetic association to biology is not always straightforward because an association between a genetic variant and a trait may not be informative with respect to the target gene (Gallagher et al., 2018). In this study, we implemented univariate and multivariate linear mixed models for starch and protein content using sorghum association panel, and identified candidate genes within significantly associated loci. We also performed candidate gene network analysis using protein-protein interaction and studied expression profile of candidate genes and their interactors.

Materials and Methods

Plant material

A panel of approximately 400 diverse sorghum accessions was planted in randomized complete block design with two replications in 2013, 2014, and 2017 field seasons at the Clemson University Pee Dee Research and Education Center in Florence, SC. This diversity panel, with over 80% of the accessions from the original United States sorghum association panel (SAP) developed by Casa et al. (2008), will be referred to as SAP. The details on experimental field design and agronomic practices have been described in details in Boyles et al. (2016) and Sapkota et al. (2020). Succinctly, the experiments were planted in a two row plots each 6.1 m long, separated by row spacing of 0.762 m with an approximate density of 130,000 plants ha^{-1}. Fields were irrigated only when signs of drought stress was seen across the field. Primary panicle of three plants selected from each plot was harvested at physiological maturity. The plants from beginning and end of the row were excluded to account for border effect. Panicles were air dried to a constant moisture (10-12%) and threshed. A 25g of cleaned and homogenized subsample of grain ground to 1-mm particle size with a CT 193 Cyclotec Sample Mill (FOSS North America) was used in near infrared spectroscopy (NIRS) for compositional analysis.

Phenotypic data

A DA 7250^{TM} NIR analyzer (Perten Instruments) was used for compositional analysis. The predicted phenotypic values were obtained from the calibrated curves for spectral measurements of ground grain samples. The calibration curve was built using wet chemistry values from a subset of samples. The wet chemistry was performed by Dairyland Laboratories, Inc. (Arcadia, WI) and the Quality Assurance Laboratory at Murphy-Brown, LLC (Warsaw, NC). The details on the prediction curves and wet chemistry can be found in Boyles et al. (2017).

The phenotypic values were fitted into a linear mixed model analysis using lme4 package in R (Bates et al., 2015; R Core Team, 2019). The following mixed model equation was fit:

$$y_{ijk} \sim G_i + Y_j + G_i \times Y_j + Y_j \times R_k + \epsilon_{ijk} \tag{1}$$

where y_{ijk} represents the phenotypic value for the combination of genotype i, year j, and replication k; G_i, Y_j, $G_i \times Y_j$, and $Y_j \times R_k$ are random effects of genotype, year, genotype-by-year, and replication-by-year, respectively; and ϵ_{ijk} is the random effect of residuals, with $N(0, \sigma^2_\epsilon)$. Best linear unbiased predictors (BLUPs) for the traits were calculated as the random effects of genotypes in the model. Variance components for genotype (G), environment/year (Y), and genotype $\times$ environment interactions were used to calculate the broad sense heritability:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_{G\times Y}}{Y} + \frac{\sigma^2_\epsilon}{YR}} \tag{2}$$

Genotypic data

The population was genetically characterized using genotyping-by-sequencing (Morris et al., 2013; Boyles et al., 2016). Sequenced reads were aligned to the BTx623 v3.1 reference assembly (phytozome) using Burrows-Wheeler aligner (Heng Li et al., 2010). TASSEL 5.0 pipeline was used for SNP calling, imputation and filtering (Glaubitz et al., 2014). The missing genotypes were imputed using the TASSEL plugin FILLIN (Swarts et al., 2014). Following imputation SNPs with minor allele frequency (MAF) <0.01, and sites missing in more than 30% genotypes in diversity panel were filtered. Genotypes with more than 10% of SNP sites missing were filtered. A total of 389 genotypes with 224,007 SNPs were used in the study. The SNP genotype file was converted into plink (Purcell et al., 2007) binary ped and bed format for association and linkage disequilibrium (LD) analysis.

Genome wide association analysis

Genome-wide association between SNPs and phenotypes were computed using a univariate or multivariate linear mixed model (LMM) fit with GEMMA v0.94 (Zhou et al., 2012; Zhou et al., 2014). Eigenvalues (-d) and eigenvectors (-u) from a genomic relationship matrix calculated using (VanRaden, 2008) was used to account for relatedness between individuals. P-values of each marker association tests were computed using Wald's statistics (-lmm 1). The SNPs with minor allele frequency less than 5% were filtered out during association analysis. Significance of marker association was determined using bonferroni threshold ($\frac{\alpha}{p}$), where $\alpha = 0.05$ and p = total number of markers.

Gene network and expression analysis

Candidate genes in LD with significantly associated SNPs were identified using annotations for BTx623 v3.1.1 (phytozome). Python codes were used to isolate associated candidate genes from annotation file and to convert gene names from *Sobic* to *Sb* gene format. Once converted, candidate genes from associated region were used to identify their high confidence (0.6) first interactors (neighbors) using sorghum protein interaction data from STRING v11.0 (www.string-db.org). Gene expression results from Olson et al. (2014) was used to examine the gene expression pattern of genes and interactors for various tissue types.

Results

Phenotypic analysis

We fit a linear mixed model to account for random effects due to environment. The genotypic effects acccounted for about 30% and 45% of total variance in protein and starch, respectively (Supplementary Table S1). The environmental variable (Year) didn't have any effect on starch, whereas year effects amounted to 17% of total variance for protein. Genotype $\times$ environment effect was slightly higher for starch (14%) than for protein (10%). The broad sense heritability was high for both protein (0.75) and starch (0.8). Both protein and starch were normally distributed, and were strongly negatively correlated to each other (Fig 1).

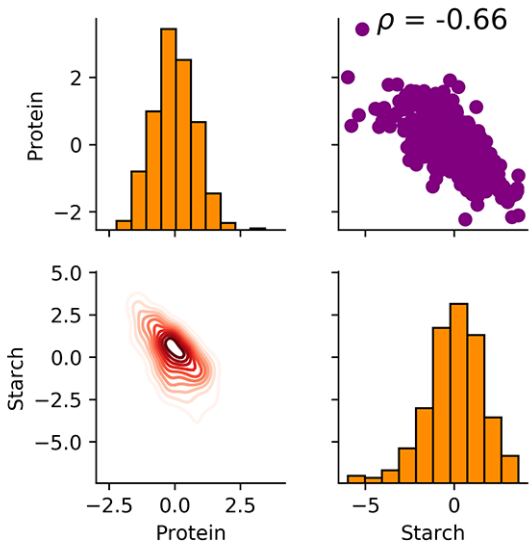


Figure 1. Distribution of the adjusted phenotypic mean (BLUPs). Right and left of the diagonal shows scatterplot and density plot, respectively, and the diagonal shows histogram. ρ=pearson correlation coefficient.

Association mapping

We filtered the SNPs with minor allele frequency less than 5% to avoid false positives. There were no SNPs that were significantly associated with protein content (Fig 2a). Starch, on the other hand, had four SNPs (S8_51720767, S8_51721062, S8_51721065, and S8_51726098) in chromosome 8 that were above the significance threshold (Fig 2b). Since starch and protein were strongly correlated, we fit a multivariate (MV) LMM to identify any other associated regions. We found two SNPs, S4_60623675 and S4_63400335, in chromosome 4 that showed significant association for MV model (Fig 2c). The SNPs on chromosome 8 that were significant for starch were also significant for the MV-LMM. Additionally, two more SNPs (S8_51715166 and S8_51719704) on chromosome 8 nearby the other associated SNPs were significantly associated in the MV analysis. All six significant SNPs in chromosome 8 were in strong linkage disequilibrium (LD) with each other (Fig 3). We also fit a univariate LMM for starch with protein as a covariate (say, *StaCovPrt*) to compare with results from MV-LMM for starch and protein. All six chromosome 8 SNPs and the chromosome 4, SNP S4_60623675, significant for MV-LMM were also significantly associated in the *StaCovPrt* model (Fig 2d). Additionally, three SNPs near 64 Mb on chromosome 4 (S4_64019577, S4_64019590 and S4_64019619) were also found to be significantly associated for the *StaCovPrt*, while the chromosome 4 SNP (S4_63400335) from MV-LMM didn't show significant association in *StaCovPrt*. The SNPs in chromosome 4 didn't show strong LD with the neighboring SNPs (Supplementary Table S2).

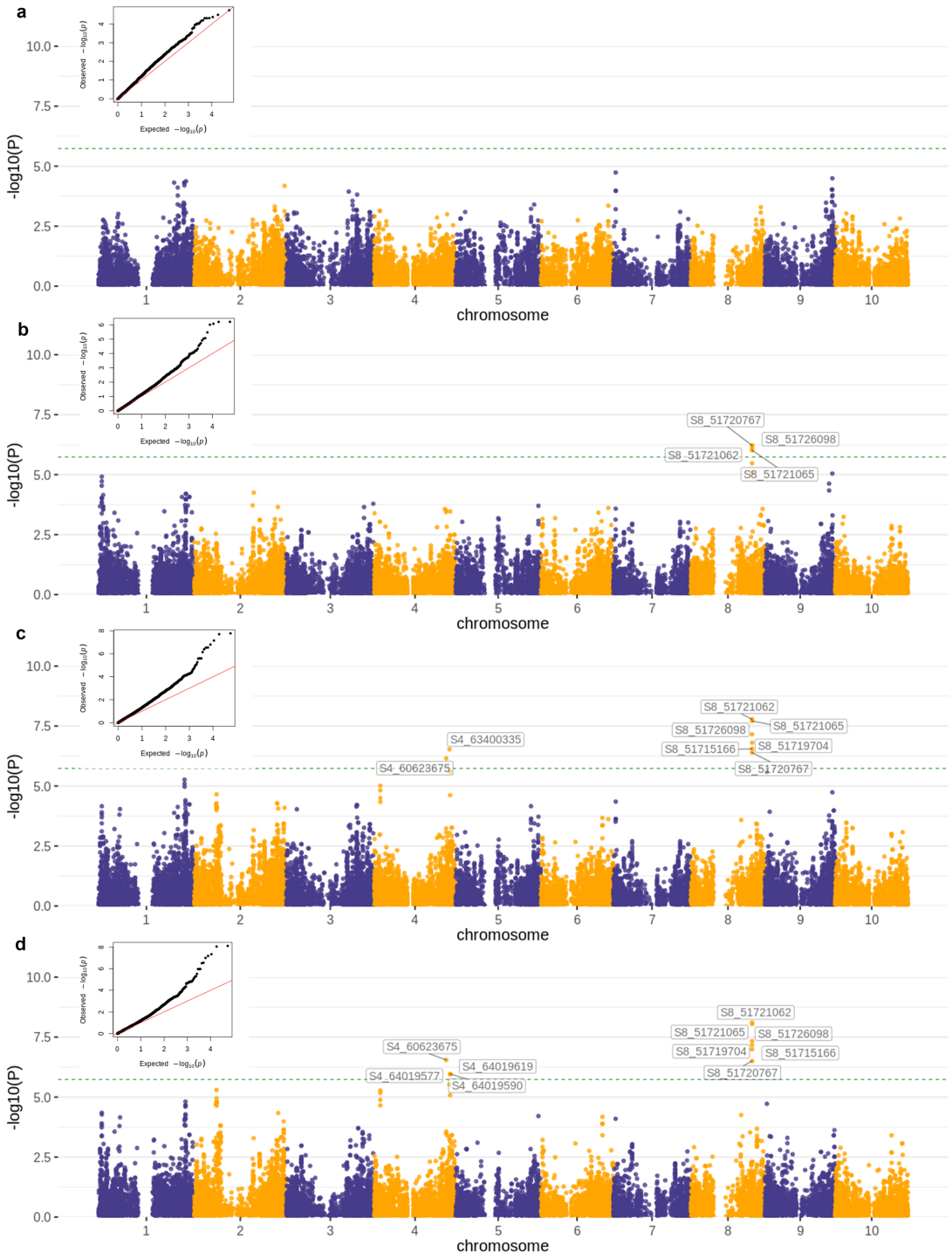


Figure 2. Manhattan plot showing genome-wide association using linear mixed model (LMM). Subfigures show univariate LMM for **a**. protein and **b**. starch, and multivariate LMM for **c**. starch and protein. Subfigure **d**. shows univariate LMM for starch with protein as covariate. Horizontal dashed green line represents Bonferroni-corrected significance threshold for α=0.05. Quantile-quantile plots for association analysis are presented as inset at top-left of subfigures. Significantly associated SNPs are annotated in the plots.

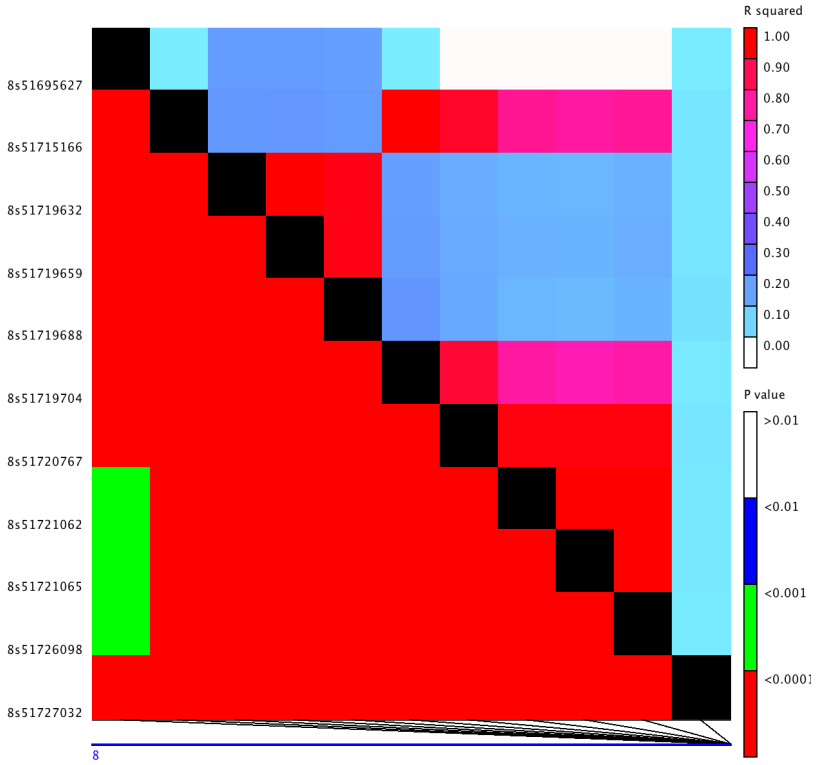


Figure 3. Linkage disequilibrium between significantly associated SNPs from chromosome 8. R-squared values to the top of the diagonal and associated p-values are at the bottom of the diagonal.

Candidate genes

We identified potential candidate genes from significantly associated SNPs based on extent of LD between SNPs in the associated regions. Since all chromosome 8 SNPs were in strong LD with each other we identified genes within the range of 51715166 bp to 51726098 bp in chromosome 8 (*Supplementary Table S2*). Since chromosome 4 SNPs showed weak LD with neighboring SNPs, we identify genes within 2 Kb of the significant SNPs as potential candidate genes.

A total of five candidate genes had significantly associated SNPs that were within or in proximity of those genes (Table 1). All SNPs except one (S4_63400335) were localized within the mRNA of the associated genes. One SNP, S8_51715166, was located in coding sequences (CDS) of a gene encoding CASP like protein (*Sobic.008G111500*), whereas, the SNPs S8_51719704 and S8_51726098 were situated within the CDS of a heat shock protein (HSP90-6; *Sobic.008G111600*). Three signficant SNPs from ∼64 Mb region of chromosome 4 were located in the 3'UTR region of a Ring-H2 finger protein, *Sobic.004G301300*. One SNP, S4_60623675, was localized in the 5'UTR region of fatty acid desaturase (FAD2) gene, *Sobic.004G260800*.

Table 1. Potential candidate genes from the significantly associated regions.

Gene	Name*	Chr	Start	End	Associated_SNPs
Sobic.004G260700	Uncharacterized protein	4	60621941	60622538	S4_60623675
Sobic.004G260800	FAD2	4	60623621	60625764	S4_60623675
Sobic.004G301300	RING-H2 finger protein	4	64018712	64019678	3 SNPs
Sobic.008G111500	CASP-like protein 8	8	51714673	51715254	S8_51715166
Sobic.008G111600	HSP-90-6	8	51719209	51726960	5 SNPs

* Gene names based on annotated homologous maize genes..

Gene network and expression

We used the string-db (or whichever) to identify high confidence (0.6) first interactors of candidate genes. The two genes in chromosome 4 had a total of 10 first interactors (Fig 4). HSP90-6 (*Sobic.008G111600*), the only chromosome 8 gene with first neighbors, had a total of 142 interactors. The gene interaction network for both sets of genes had higher number of protein-protein interaction (PPI) than expected (PPI enrichment p-value $<1e^{-16}$). Gene interaction networks from chromosome 4 and chromosome 8 were significantly enriched (FDR<0.001) for three and six biochemical pathways, respectively (Supplementary Table S3). The chromosome 4 genes were enriched for biosynthesis of unsaturated fatty acids, fatty acid metabolism and nitrogen metabolism pathways, whereas, chromosome 8 genes were enriched for protein processing, protein export, spliceosome, endocytosis, RNA degradation, and plant-pathogen interaction. Figure 5 shows expression atlas of chromosome 4 and chromosome 8 group of genes and interactors across various tissue types. While clear clustering of genes with differential expression across reproductive and vegetative tissues was not seen, different clusters of genes showed varying degree of transcript abundance across tissue types.

The FAD2 gene (*Sobic.004G260800*) in chromosome 4 interacted with another fatty acid desaturase gene (DES2, *Sobic.004G260600*) which is located ∼570 Kb upstream from the FAD2 gene. Both FAD2 and DES2 strongly interact with three nitrate reductase (NADH) genes, two of which located in chromosome 4 at ∼55 Mb (*Sobic.004G196101*) and ∼65 Mb (*Sobic.004G312500*) while the other is located around 58-59 Mb in chromosome 7 (*Sobic.007G153900*) (Figure 4). The FAD2 gene was highly expressed in flower, embryo and shoot, whereas, its interactor DES2 had higher expression in root tissues (Fig 6). The heat shock proteins are known to be molecular chaperones primarily involved in drought and stress response but can also be involved in other molecular processes during plant development (Yu et al., 1998; Khan et al., 2019). The gene expression results showed HSP-90 (*Sobic.008G111600*) to be highly expressed in floral meristem, plant embryo and vegetative meristem compared to root, shoot, and flower tissues (Fig 6). The interactors of HSP90-6 included several

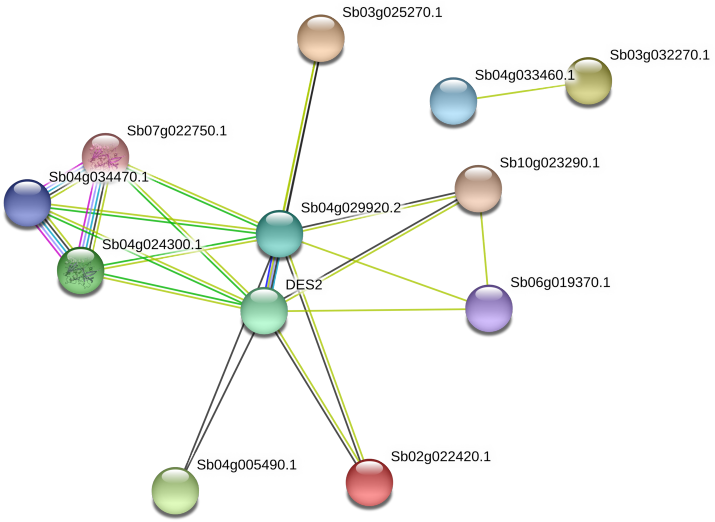


Figure 4. Network of candidates genes (*Sobic.004G260800.1* and *Sobic.004G033460.1*) and interactors from associated chromosome 4 SNPs.

HSP70 genes which were also highly expressed in reproductive tissues compared to root and shoot tissues (Fig 6).

Discussion

Despite being two of the most studied grain quality phenotypes, large proportion of genetic variance in starch and protein remains unexplained. In this study, we aimed to identify genetic loci associated with starch and protein content in sorghum grain. We observed strong genotypic effect and some genotype $\times$ environment effect for starch and protein in our population (Supplementary Table S1). Previous studies have also reported high heritability for starch and protein in different populations (Rami et al., 1998; Murray et al., 2008; Rhodes et al., 2017).

In our genome-wide association study, we used only the random genetic effects (BLUPs) to identify marker trait association for strictly genetic effects. The lack of genetic association for protein content could be due to smaller genetic effects and larger environmental effects on this trait. Starch, with no environmental effect and larger genotypic effects, had genetic variants significantly associated with the phenotype. Five SNPs (in strong LD with each other) from a single locus that encodes for a heat shock protein (HSP) 90 (*Sobic.008G111600*), with two SNPs located on the coding sequence, showed significant association. This loci was not identified during association mapping of starch in previous studies using this population (Rhodes et al., 2017; Boyles et al., 2017). HSPs are common group of

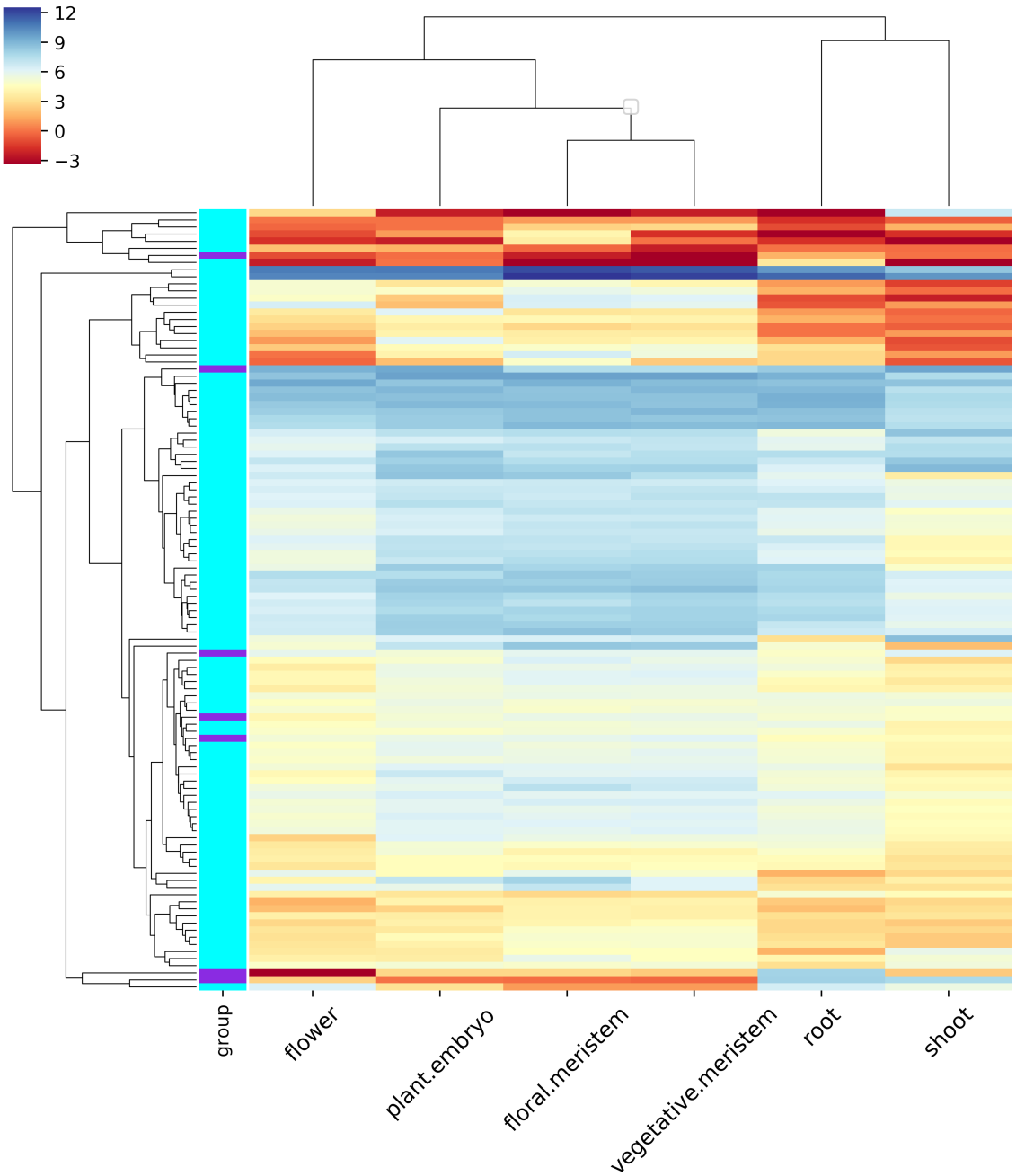


Figure 5. Heatmap showing gene expression analysis of interactors of candidate genes. The row colors represents chromosome 'group': purple and cyan represent chromosome 4 and 8 related genes, respectively.

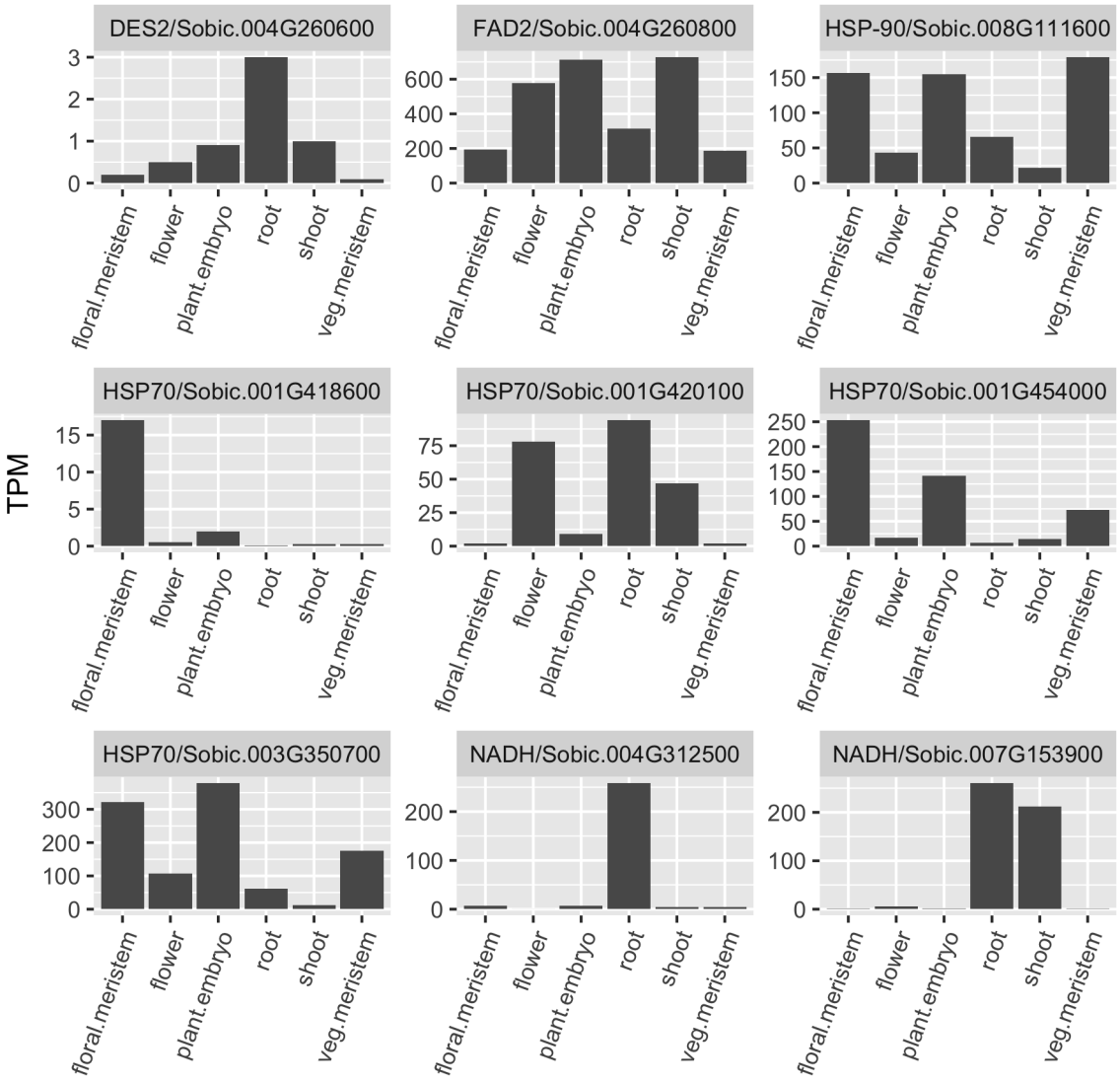


Figure 6. Gene expression of candidate genes and some of their interactors. X-axis represents various tissue types, and y-axis shows relative expression (TPM: transcript per million). HSP: heat shock protein, FAD/DES: fatty acid desaturase, NADH: nitrate reductase. Expression data obtained from Olson et al. (2014).

protein found in eukaryotes and function as molecular chaperones that help in refolding proteins denatured by heat and keep them from aggregating (Vierling, 1991; Boston et al., 1996). Two distinct members of the Hsp70 family of stress-related protein were localized in the maize amyloplast and form transient complexes with starch synthase 1 (SSI) and other stromal enzymes (Yu et al., 1998). In Japanese sake-brewing rice rich in starch content, the HSP70 protein was highly abundant in amyloplast compared to cytosol and its concentration was elevated during the later stages of grain development (Kamara et al., 2009). The HSP90 gene in this study was seen to be strongly interacting with numerous molecular chaperones including HSP70. Since the HSP90 and interacting HSP70 proteins showed higher expression in the embryo and floral meristem compared to root and shoot tissues, they are likely important candidates responsible for protein processing and export during grain filling in sorghum.

For complex traits, understanding if a genetic variant affects multiple phenotypes simultaneously (pleiotropy) or affects one phenotype through affecting another phenotype is one of the major challenge (Yang et al., 2012). Starch and protein constitute most of the grain composition and display a strong negative correlation. We identified additional marker trait associations when: 1) starch and protein were fit as dependent variables in a multivariate mixed model, or 2) when protein was fit as independent variable for a model with starch as dependent variable. This approach helped us identify an important variant, S4_60623675, which showed significant association in both of the above mentioned models and was located in the 5' UTR of a fatty acid desaturase gene (*Sobic.004G260800*). The fatty acid desaturase gene interacted with two more desaturase genes and three nitrate reductase genes, forming a network that is highly enriched for biochemical pathways for fatty acid and nitrogen. One of the interacting desaturase genes (DES2, *Sobic.004G260600*) is involved in the biosynthesis of aliphatic side chain of sorgoleone by converting palmitoleic acid to hexadecadienoic acid (Pan et al., 2007). Sorgoleone is a phytotoxic secondary metabolite that plays a direct role in allelopathic interactions. Sorgoleone is known to inhibit photosynthesis, but the relationship between Sorgoleone biosynthesis pathway and seed development is unclear (Einhellig et al., 1993). Rhodes et al. (2017) have previously reported significantly associated variants for protein and fat around 57-58 Mb of chromosome 4 which is ∼2-3 Mb from our associated SNPs. The high expression of HSP90 gene in plant embryo combined with a candidate SNPs in the fatty acid desaturase gene identified using MV-LMM hints at possible connection between the biochemical pathways for starch, protein and fat content during grain development. The enrichment of biochemical pathways involving these genes and their high expression in reproductive tissues warrants further characterization and functional

analysis of these candidates.

In conclusion, we were able to identify a previously uncharacterized genomic region associated with starch content using univariate LMM. The genomic region harbored a heat shock protein which shows strong protein-protein interaction and its network is enriched for several biochemical pathways. Additionally, we also showed that use of MV-LMM for correlated traits can help identify additional genomic regions that go undetected with the univariate GWAS of single traits. The candidates of this study might be involved in intricate metabolic pathway and represent possible pleiotropic targets for source-sink activities during grain filling.

Supporting Information

Supplementary information is included in Appendix C.

Data availability

The codes and data used in the study are available at github.com/sirjansapkota/StarchProtein.

References

Aschard, Hugues, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft (2014). "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies". In: *The American Journal of Human Genetics* 94.5, pp. 662–676.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Boston, Rebecca S, Paul V Viitanen, and Elizabeth Vierling (1996). "Molecular chaperones and protein folding in plants". In: *Post-transcriptional control of gene expression in plants*. Springer, pp. 191–222.

Boyles, Richard E, Elizabeth A Cooper, Matthew T Myers, Zachary Brenton, Bradley L Rauh, Geoffrey P Morris, and Stephen Kresovich (2016). "Genome-wide association studies of grain yield components in diverse sorghum germplasm". In: *The plant genome* 9.2.

Boyles, Richard E, Brian K Pfeiffer, Elizabeth A Cooper, Bradley L Rauh, Kelsey J Zielinski, Matthew T Myers, Zachary Brenton, William L Rooney, and Stephen Kresovich (2017). "Genetic dissection of sorghum grain quality traits using diverse and segregating populations". In: *Theoretical and applied genetics* 130.4, pp. 697–716.

Carlson, Maryn O, Gracia Montilla-Bascon, Owen A Hoekenga, Nicholas A Tinker, Jesse Poland, Matheus Baseggio, Mark E Sorrells, Jean-Luc Jannink, Michael A Gore, and Trevor H Yeats (2019). "Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (Avena sativa L.)" In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2963–2975.

Casa, Alexandra M, Gael Pressoir, Patrick J Brown, Sharon E Mitchell, William L Rooney, Mitchell R Tuinstra, Cleve D Franks, and Stephen Kresovich (2008). "Community resources and strategies for association mapping in sorghum". In: *Crop science* 48.1, pp. 30–40.

Cook, Jason P, Michael D McMullen, James B Holland, Feng Tian, Peter Bradbury, Jeffrey Ross-Ibarra, Edward S Buckler, and Sherry A Flint-Garcia (2012). "Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels". In: *Plant physiology* 158.2, pp. 824–834.

Einhellig, Frank A, James A Rasmussen, Angela M Hejl, and Itamar F Souza (1993). "Effects of root exudate sorgoleone on photosynthesis". In: *Journal of chemical ecology* 19.2, pp. 369–375.

Gaire, Rupesh, Mao Huang, Clay Sneller, Carl Griffey, Gina Brown-Guedira, and Mohsen Mohammadi (2019). "Association Analysis of Baking and Milling Quality Traits in an Elite Soft Red Winter Wheat Population". In: *Crop Science* 59.3, pp. 1085–1094.

Gallagher, Michael D and Alice S Chen-Plotkin (2018). "The post-GWAS era: from association to function". In: *The American Journal of Human Genetics* 102.5, pp. 717–730.

Gieger, Christian, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé De Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H-Erich Wichmann, Klaus M Weinberger, Jerzy Adamski, et al. (2008). "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum". In: *PLoS genetics* 4.11.

Glaubitz, Jeffrey C, Terry M Casstevens, Fei Lu, James Harriman, Robert J Elshire, Qi Sun, and Edward S Buckler (2014). "TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline". In: *PloS one* 9.2, e90346.

Kamara, Joseph S, Miki Hoshino, Yuki Satoh, Nasrin Nayar, Motoko Takaoka, Tsuneo Sasanuma, and Toshinori Abe (2009). "Japanese sake-brewing rice cultivars show high levels of globulin-like protein and a chloroplast stromal HSP70". In: *Crop science* 49.6, pp. 2198–2206.

Khan, Abid, Muhammad Ali, Abdul Mateen Khattak, Wen-Xian Gai, Huai-Xia Zhang, Ai-Min Wei, Zhen-Hui Gong, et al. (2019). "Heat Shock Proteins: Dynamic Biomolecules to Counter Plant Biotic and Abiotic Stresses". In: *International journal of molecular sciences* 20.21, p. 5321.

Korte, Arthur and Ashley Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1, p. 29.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). "A mixed-model approach for genome-wide association studies of correlated traits in structured populations". In: *Nature genetics* 44.9, p. 1066.

Li, Heng and Richard Durbin (2010). "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5, pp. 589–595.

Mace, Emma S, Shuaishuai Tai, Edward K Gilding, Yanhong Li, Peter J Prentis, Lianle Bian, Bradley C Campbell, Wushu Hu, David J Innes, Xuelian Han, et al. (2013). "Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum". In: *Nature communications* 4, p. 2320.

Morris, Geoffrey P, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. (2013). "Population genomic and genome-wide association studies of agroclimatic traits in sorghum". In: *Proceedings of the National Academy of Sciences* 110.2, pp. 453–458.

Murray, Seth C, Arun Sharma, William L Rooney, Patricia E Klein, John E Mullet, Sharon E Mitchell, and Stephen Kresovich (2008). "Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates". In: *Crop Science* 48.6, pp. 2165–2179.

Myles, Sean, Jason Peiffer, Patrick J Brown, Elhan S Ersoz, Zhiwu Zhang, Denise E Costich, and Edward S Buckler (2009). "Association mapping: critical considerations shift from genotyping to experimental design". In: *The Plant Cell* 21.8, pp. 2194–2202.

Olson, Andrew, Robert R Klein, Diana V Dugas, Zhenyuan Lu, Michael Regulski, Patricia E Klein, and Doreen Ware (2014). "Expanding and vetting Sorghum bicolor gene annotations through transcriptome and methylome sequencing". In: *The Plant Genome* 7.2.

Pan, Zhiqiang, Agnes M Rimando, Scott R Baerson, Mark Fishbein, and Stephen O Duke (2007). "Functional characterization of desaturases involved in the formation of the terminal double bond of an unusual 16: 3Δ9, 12, 15 fatty acid isolated from Sorghum bicolor root hairs". In: *Journal of Biological Chemistry* 282.7, pp. 4326–4335.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3, pp. 559–575.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rami, J-F, Philippe Dufour, Gilles Trouche, Geneviève Fliedel, Christian Mestres, Fabrice Davrieux, P Blanchard, and Perla Hamon (1998). "Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (Sorghum bicolor L. Moench)". In: *Theoretical and applied genetics* 97.4, pp. 605–616.

Reif, Jochen C, Manje Gowda, Hans P Maurer, CFH Longin, Viktor Korzun, Erhard Ebmeyer, Reiner Bothe, Christof Pietsch, and Tobias Würschum (2011). "Association mapping for quality traits in soft winter wheat". In: *Theoretical and Applied Genetics* 122.5, pp. 961–970.

Rhodes, Davina H, Leo Hoffmann, William L Rooney, Thomas J Herald, Scott Bean, Richard Boyles, Zachary W Brenton, and Stephen Kresovich (2017). "Genetic architecture of kernel composition in global sorghum germplasm". In: *BMC genomics* 18.1, p. 15.

Rice, Brian R, Samuel B Fernandes, and Alexander E Lipka (2020). "Multi-Trait Genome-wide Association Studies Reveal Loci Associated with Maize Inflorescence and Leaf Architecture". In: *Plant and Cell Physiology*.

Sapkota, Sirjan, Rick Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich (Jan. 2020). "Impact of sorghum racial structure and diversity on genomic prediction of grain yield components". In: *Crop Science*. DOI: `10.1002/csc2.20060`.

Sukumaran, Sivakumar, Wenwen Xiang, Scott R Bean, Jeffrey F Pedersen, Stephen Kresovich, Mitchell R Tuinstra, Tesfaye T Tesso, Martha T Hamblin, and Jianming Yu (2012). "Association mapping for grain quality in a diverse sorghum collection". In: *The Plant Genome* 5.3, pp. 126–135.

Swarts, Kelly, Huihui Li, J Alberto Romero Navarro, Dong An, Maria Cinta Romay, Sarah Hearne, Charlotte Acharya, Jeffrey C Glaubitz, Sharon Mitchell, Robert J Elshire, et al. (2014). "Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants". In: *The Plant Genome* 7.3.

Thoen, Manus PM, Nelson H Davila Olivas, Karen J Kloth, Silvia Coolen, Ping-Ping Huang, Mark GM Aarts, Johanna A Bac-Molenaar, Jaap Bakker, Harro J Bouwmeester, Colette Broekgaarden,

et al. (2017). “Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping”. In: *New Phytologist* 213.3, pp. 1346–1362.

VanRaden, Paul M (2008). “Efficient methods to compute genomic predictions”. In: *Journal of dairy science* 91.11, pp. 4414–4423.

Vierling, Elizabeth (1991). “The roles of heat shock proteins in plants”. In: *Annual review of plant biology* 42.1, pp. 579–620.

Wang, Xiaoqian, Yunlong Pang, Jian Zhang, Zhichao Wu, Kai Chen, Jauhar Ali, Guoyou Ye, Jianlong Xu, and Zhikang Li (2017). “Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content”. In: *Scientific reports* 7.1, pp. 1–10.

Wilson, Larissa M, Sherry R Whitt, Ana M Ibáñez, Torbert R Rocheford, Major M Goodman, and Edward S Buckler (2004). “Dissection of maize kernel composition and starch production by candidate gene association”. In: *The Plant Cell* 16.10, pp. 2719–2733.

Yang, Qiong and Yuanjia Wang (2012). “Methods for analyzing multivariate phenotypes in genetic association studies”. In: *Journal of probability and statistics* 2012.

Yang, Qiong, Hongsheng Wu, Chao-Yu Guo, and Caroline S Fox (2010). “Analyze multivariate phenotypes in genetic association studies by combining univariate association tests”. In: *Genetic epidemiology* 34.5, pp. 444–454.

Yu, Ying, Helen He Mu, Chen Mu-Forster, and Bruce P Wasserman (1998). “Polypeptides of the maize amyloplast stroma: stromal localization of starch-biosynthetic enzymes and identification of an 81-kilodalton amyloplast stromal heat-shock cognate”. In: *Plant physiology* 116.4, pp. 1451–1460.

Zhao, Keyan, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. (2011). “Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa”. In: *Nature communications* 2.1, pp. 1–10.

Zhou, Xiang and Matthew Stephens (2012). “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7, pp. 821–824.

Zhou, Xiang and Matthew Stephens (2014). “Efficient multivariate linear mixed model algorithms for genome-wide association studies”. In: *Nature methods* 11.4, p. 407.

Zhu, Fan (2014). “Structure, physicochemical properties, modifications, and uses of sorghum starch”. In: *Comprehensive Reviews in Food Science and Food Safety* 13.4, pp. 597–610.

Chapter 5

Summary and Future Directions

Crop breeding has long benefited and will continue to benefit from the understanding and application of population and quantitative genetics. In my doctoral research, I studied genetic basis of quantitative inheritance for grain yield and quality and applied computational methods for genomics-assisted breeding in sorghum. The three research chapters (Chapter 2, 3 and 4) represented the three objectives that were pursued during the course of my doctoral degree.

The motivation behind Chapter 2 was to understand how population structure and genetic diversity within subpopulation affects genomic prediction accuracy in sorghum. While impact of population structure on genomic prediction had been studied previously in maize and rice (Guo et al., 2014), no such study in sorghum existed prior to our investigation. In Sapkota et al. (2020), we showed having similar population structure between training and testing populations impacts prediction accuracy positively. This observation is similar to observations in maize and rice diversity panels, however, the large difference in prediction accuracy between within subpopulation and across subpopulation predictions was much more subtle in sorghum than it was in maize and rice. As racial structure is strongly tied to population structure, the large contribution (30-50%) of races to covariance for panicle and yield traits in stratified sampling method wasn't surprising. However, one take way based on within and across race predictions can be that genetic diversity and shared alleles can be as important as genomic relationship in diverse and structured populations. Sorghum breeding has largely relied on phenotypic selection, and as a result strong biases towards races with favorable panicle architecture have justifiably existed. During the pre-breeding exercises where expansion of genetic and phenotypic diversity is important, breeder's biases towards early

generation selection based on phenotypic appearance could, however, limit genetic gain by limiting genetic variance. Our results from this investigation suggests increasing genetic diversity by including not only the US sorghum breeding ideotypes but also germplasm that allow exploiting racial differences would be beneficial. To that objective, including more guinea and bicolor accessions as training individuals for screening germplasm for pre-breeding would be advantageous because our results show that accessions in these races boost prediction accuracy when used as training or testing population. A study of several biparental populations or multiparental recombinant inbred populations in sorghum to examine cross validation strategies such as ours would be valuable for insights into effects of family structure in prediction accuracy.

Two motivating factors were driving the research in Chapter 3: a) phenotyping using near-infrared spectroscopy is labor-intensive, generally destructive, and time limiting; and b) finding and testing models to incorporate multi-dimensional data structure of breeding programs. Our results show high accuracy of prediction for grain compositional traits, and that prediction accuracy can be increased substantially by accounting for genotype $\times$ environment variance and trait correlations. The ability to predict line performance in an unobserved environment is of great importance to breeding programs, and results show high accuracy for predicting whole environments using Bayesian multi-output regressor stacking (BMORS) model. This prediction model needs to be tested on more populations and traits, especially for grain yield components because traits such as grain number and grain size that are strongly negatively correlated can benefit from models that can exploit the trade-offs during selection.

Several association studies have previously been conducted for starch and protein content using the US sorghum association panel (Boyles et al., 2017; Rhodes et al., 2017; Sukumaran et al., 2012). In our Chapter 4, we conducted association analysis using univariate and multivariate linear mixed models (LMM) for best linear unbiased predictions (BLUPs) of starch and protein. For univariate LMM, protein didn't show any significant marker trait associations, but we were able to identify a previously uncharacterized genomic region in chromosome 8 that is associated with starch content. For multivariate LMM using starch and protein, we were able to identify additional significantly associated variants in chromosome 4. We identified two potential candidate genes, one from chromosome 8 (HSP90) and another from chromosome 4 (FAD2) that showed strong protein-protein interactions with several first neighbors. Since these genes showed relatively high expression in reproductive tissues and were also enriched for biochemical pathways, they might

be important candidates involved in biochemical processes during grain filling. Further analysis for these associated regions would involve identifying different haplotypic groups and check for phenotypic variation associated with haplotypic variation. Since two SNPs were situated within coding sequences of HSP90, study of variant effect on gene function and protein confirmation will be another follow up analysis.

Lastly, genomic selection has brought about paradigm shift in plant breeding. While this study is a puzzle-solving science, I believe it will provide information that will act as stepping stone in streamlining application of genomic selection in sorghum breeding. In bigger picture, data science has changed the intensity of scientific progress. However, it will never replace the need for sound science which I believe stands tall on the shoulders of plausible hypothesis set forth by detailed scientific process.

References

Boyles, Richard E, Brian K Pfeiffer, Elizabeth A Cooper, Bradley L Rauh, Kelsey J Zielinski, Matthew T Myers, Zachary Brenton, William L Rooney, and Stephen Kresovich (2017). "Genetic dissection of sorghum grain quality traits using diverse and segregating populations". In: *Theoretical and applied genetics* 130.4, pp. 697–716.

Guo, Zhigang, Dominic M Tucker, Christopher J Basten, Harish Gandhi, Elhan Ersoz, Baohong Guo, Zhanyou Xu, Daolong Wang, and Gilles Gay (2014). "The impact of population structure on genomic prediction in stratified populations". In: *Theoretical and applied genetics* 127.3, pp. 749–762.

Rhodes, Davina H, Leo Hoffmann, William L Rooney, Thomas J Herald, Scott Bean, Richard Boyles, Zachary W Brenton, and Stephen Kresovich (2017). "Genetic architecture of kernel composition in global sorghum germplasm". In: *BMC genomics* 18.1, p. 15.

Sapkota, Sirjan, Richard Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich (2020). "Impact of sorghum racial structure and diversity on genomic prediction of grain yield components". In: *Crop Science*.

Sukumaran, Sivakumar, Wenwen Xiang, Scott R Bean, Jeffrey F Pedersen, Stephen Kresovich, Mitchell R Tuinstra, Tesfaye T Tesso, Martha T Hamblin, and Jianming Yu (2012). "Association mapping for grain quality in a diverse sorghum collection". In: *The Plant Genome* 5.3, pp. 126–135.

Appendices

Appendix A Supplementary File Chapter 2

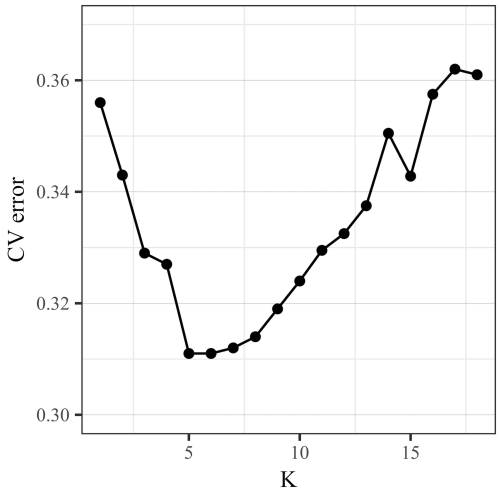
SUPPLEMENTARY FIGURES AND TABLES

Impact of Sorghum Racial Structure and Diversity on Genomic Prediction of Grain Yield Components

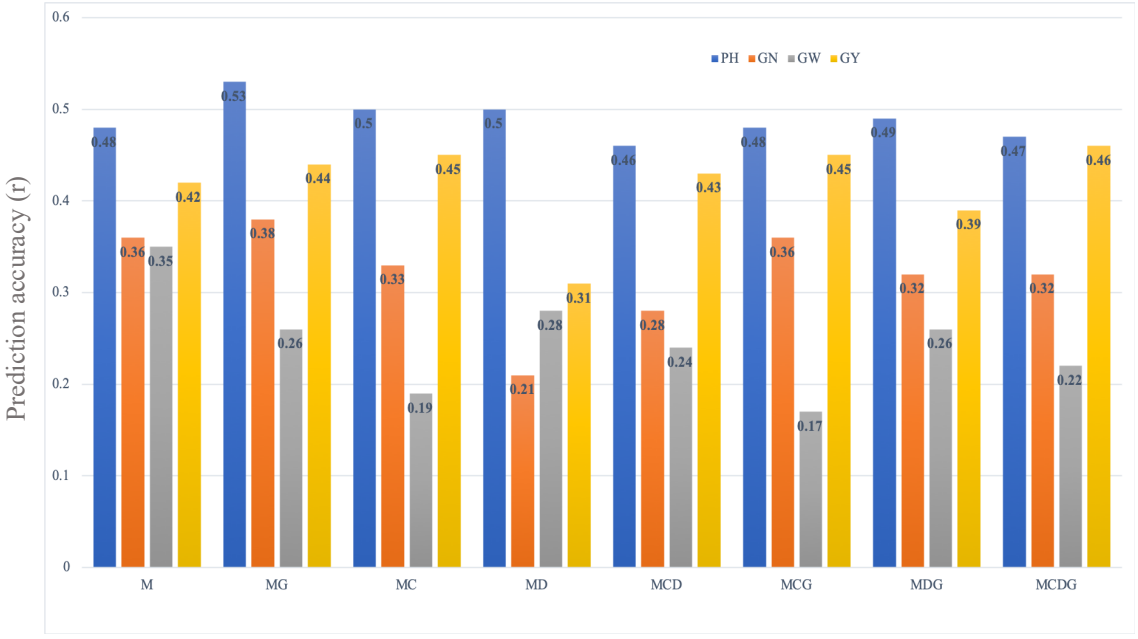
Sirjan Sapkota,* Richard Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich

Affiliations: S. Sapkota, Z. Brenton, M. Myers and S. Kresovich, Advanced Plant Technology Program, Clemson University, Clemson, SC 29634; S. Sapkota, R. Boyles, and S. Kresovich, Department of Plant and Environmental Sciences, Clemson University, Clemson SC 29634; R. Boyles, Pee Dee Research and Education Center, Clemson University, Florence, SC 29506; E. Cooper, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223.

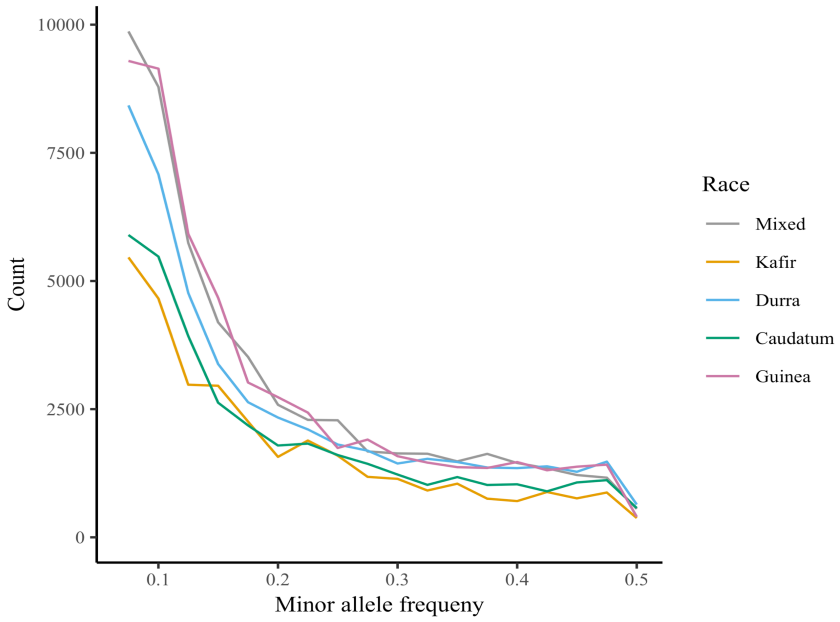
Supplementary Figure S1. Cross validation (CV) of error for number of clusters (K = 1 to 18) in admixture.



Supplementary Figure S2. Cross-validation accuracy for 36 kafir accessions using combination of one or many other races as training population. Method: We randomly sampled 36 kafir accessions and used it as validation population for the all cross-validation runs. The training population size used was always 36 individuals. A total of 36, 18, 12, and nine individuals from each training race was used when one, two, three, and four training race were used, respectively. The alphabets in x-axis represent races: M = mixed, C = caudatum, D = durra, and G = guinea. Colors in bars represent traits: GN= grain number, GW = grain weight, GY = grain yield, and PH = plant height. The respect numerical values for prediction accuracies are shown at the top of the bars.



Supplementary Figure S3. Allele frequency spectra for minor alleles across accessions within each race. The minor alleles were plotted at *binwidth* of 0.02 with total number (count) of alleles with that minor allele frequency.



Supplementary Table S1. Genetic differentiation between the races based on F_{st} (bottom and left of the diagonal) and Euclidean distance between centroids (top and right of the diagonal) of race clusters from first five principal components.

	Mixed	Kafir	Durra	Caudatum	Guinea
Mixed		41.12	38.2	30.07	36.03
Kafir	0.08		61.34	60.79	63.12
Durra	0.07	0.16		54.76	55.32
Caudatum	0.05	0.15	0.13		54.71
Guinea	0.07	0.17	0.12	0.14	

Supplementary Table S2. Mean and standard deviation of all phenotypic traits by sorghum races. BL, terminal branch length; DTA, days to anthesis; FLH, flag leaf height; GN, grain number; GW, grain weight; GY, grain yield; PH, plant height; PL, panicle length.

Race	DTA	PH	GN	GW	GY	FLH	PL	BL
		cm		g	g	cm	cm	cm
Caudatum	75 ± 8	119.74 ± 43.18	1472 ± 407	24.99 ± 4.87	47.13 ± 13.47	90.15 ± 40.91	19.93 ± 4.04	6.31 ± 1.94
Durra	78 ± 10	124.18 ± 47.59	1118 ± 362	21.53 ± 5.66	33.05 ± 12.55	87.2 ± 38.89	19.84 ± 6.8	7.56 ± 4.40
Guinea	70 ± 8	127.02 ± 49.08	1032 ± 423	25.21 ± 6.69	35.06 ± 17.08	81.97 ± 19.36	28.73 ± 6.44	9.74 ± 3.13
Kafir	71 ± 6	130.74 ± 55.9	1404 ± 409	22.07 ± 4.25	39.6 ± 11.69	95.9 ± 47.08	21.87 ± 4.07	7.37 ± 1.78
Mixed	71 ± 8	131.21 ± 50.38	1228 ± 419	21.69 ± 6.12	34.08 ± 12.34	92.68 ± 36.8	23.55 ± 6.08	8.36 ± 2.99
Total	74 ± 8.5	125.2 ± 48.2	1301 ± 433	23.23 ± 5.6	39.48 ± 14.5	90.11 ± 39.3	21.69 ± 6	7.43 ± 3.07

Supplementary Table S3. Mean estimates of variance and covariance for WR and AR prediction in CV2 method. AR, Across race; WR, within race.

	covariance		variance (predicted values)	
Trait	WR	AR	WR	AR
Terminal branch length (BL)	271.53	35.42	216.38	34.67
Days to anthesis (DTA)	5.95	1.31	6.06	1.84
Flag leaf height (FLH)	294.55	63.44	268.16	31.94
Grain number per panicle (GN)	8419.33	7350.35	8115.08	5448
1000-grain weight (GW)	8.5	2.16	6.17	1.22
Grain yield per panicle (GY)	20.31	13.14	14.68	9.14
Plant height (PH)	503.43	236.2	455.17	117.51
Panicle length (PL)	4.9	0.75	4.26	1.14

Supplementary Table S4. Mean prediction accuracies by races for WR and AR prediction in CV2 method. Higher mean prediction accuracy for each trait across all races and methods is highlighted in bold. Values represent mean ± standard deviation.

Traits	Caudatum		Durra		Guinea		Kafir		Mixed	
	WR	AR	WR	AR	WR	AR	WR	AR	WR	AR
Primary branch length	0.21 ± 0.04	0.03 ± 0.20	**0.81** ± 0.06	0.25 ± 0.28	0.23 ± 0.09	0.23 ± 0.39	0.26 ± 0.11	0.10 ± 0.24	0.39 ± 0.05	0.38 ± 0.28
Days to anthesis	**0.55** ± 0.02	0.34 ± 0.21	0.33 ± 0.06	0.00 ± 0.25	-0.07 ± 0.08	0.00 ± 0.44	0.06 ± 0.07	0.06 ± 0.25	0.33 ± 0.04	0.18 ± 0.25
Flag leaf height	0.48 ± 0.03	0.47 ± 0.19	**0.66** ± 0.03	0.50 ± 0.24	0.00 ± 0.10	0.01 ± 0.39	0.46 ± 0.05	0.42 ± 0.24	0.44 ± 0.05	0.32 ± 0.32
Grain number	0.36 ± 0.03	0.21 ± 0.18	0.33 ± 0.06	0.09 ± 0.25	0.18 ± 0.12	0.34 ± 0.41	0.13 ± 0.07	0.28 ± 0.24	0.26 ± 0.09	**0.42** ± 0.25
1000-grain weight	0.54 ± 0.03	0.29 ± 0.21	0.62 ± 0.01	0.45 ± 0.28	**0.77** ± 0.02	0.48 ± 0.31	0.65 ± 0.02	0.22 ± 0.24	0.49 ± 0.04	0.42 ± 0.20
Grain yield	0.32 ± 0.04	0.19 ± 0.18	0.43 ± 0.05	0.24 ± 0.27	**0.53** ± 0.05	0.44 ± 0.41	0.28 ± 0.06	0.43 ± 0.19	0.23 ± 0.09	0.46 ± 0.27
Plant height	0.55 ± 0.03	0.55 ± 0.16	**0.62** ± 0.03	0.54 ± 0.20	0.21 ± 0.06	0.27 ± 0.42	0.54 ± 0.04	0.51 ± 0.19	0.37 ± 0.05	0.40 ± 0.24
Panicle length	0.25 ± 0.07	0.02 ± 0.17	**0.79** ± 0.04	0.29 ± 0.27	0.02 ± 0.13	0.10 ± 0.46	-0.07 ± 0.09	-0.03 ± 0.27	0.25 ± 0.07	0.21 ± 0.29
Average	0.41 ± 0.14	0.26 ± 0.26	0.57 ± 0.18	0.30 ± 0.31	0.23 ± 0.28	0.23 ± 0.44	0.29 ± 0.24	0.25 ± 0.30	0.35 ± 0.11	0.35 ± 0.28

Appendix B Supplementary File Chapter 3

Multi-trait regressor stacking increased genomic prediction accuracy of sorghum grain composition

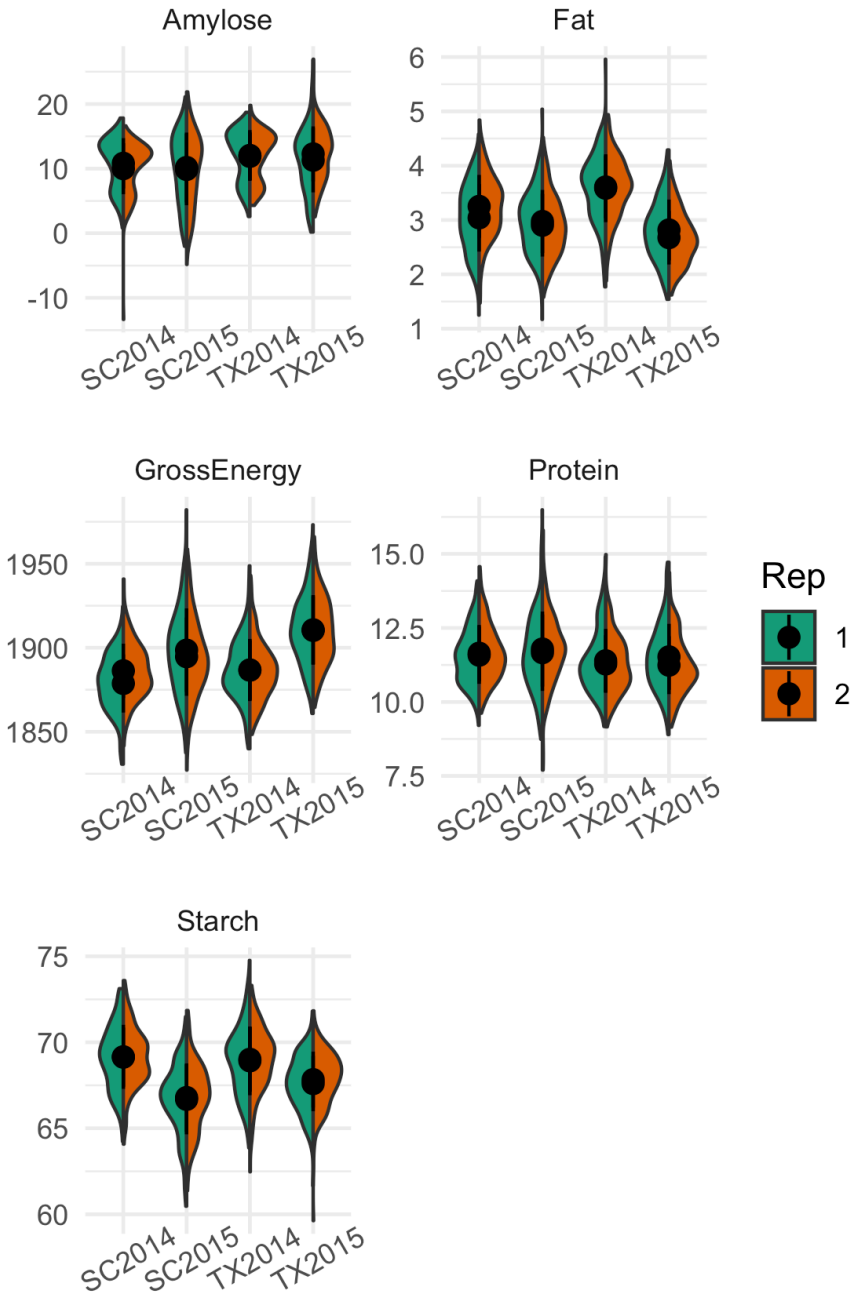
Sirjan Sapkota[1,2*], J. Lucas Boatwright[1], Kathleen E. Jordan[1], Richard E. Boyles[2,3], Stephen Kresovich[1,2]

[1]Advanced Plant Technology Program, Clemson University, Clemson, SC, USA
[2]Department of Plant and Environmental Sciences, Clemson University, Clemson, SC,USA
[3]Pee Dee Research and Education Center, Clemson University, Florence, SC, USA

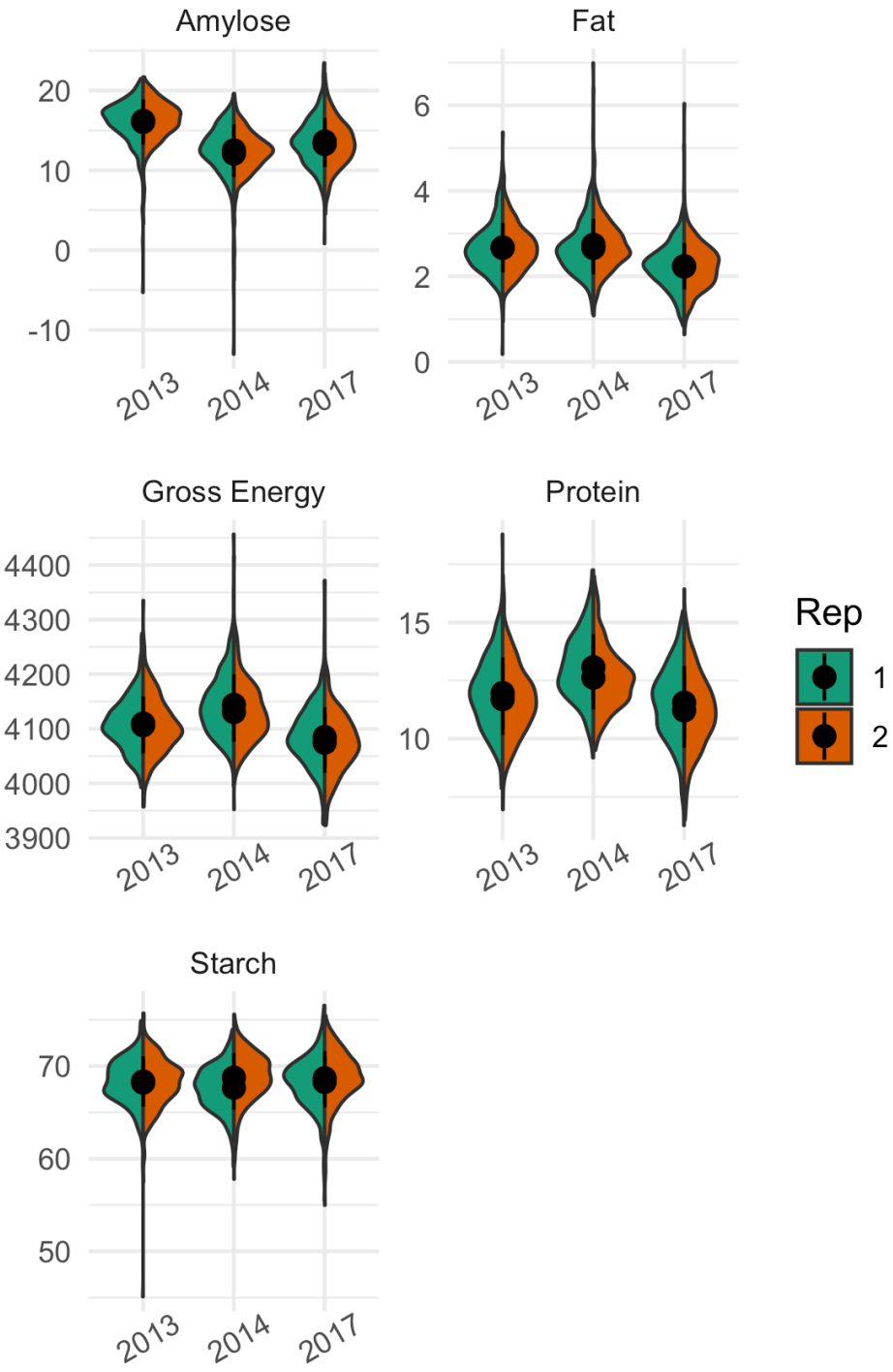
* Corresponding author: ssapkot@g.clemson.edu

Supplementary Information

S1 Fig. **Phenotypic distribution of grain composition traits in the RILs**. In the x-axes, SC: South Carolina, TX: Texas, numbers represent years. Values are percentage dry basis for protein, fat and starch; gross energy is in KCal/lb; and amylose is in percent of starch.

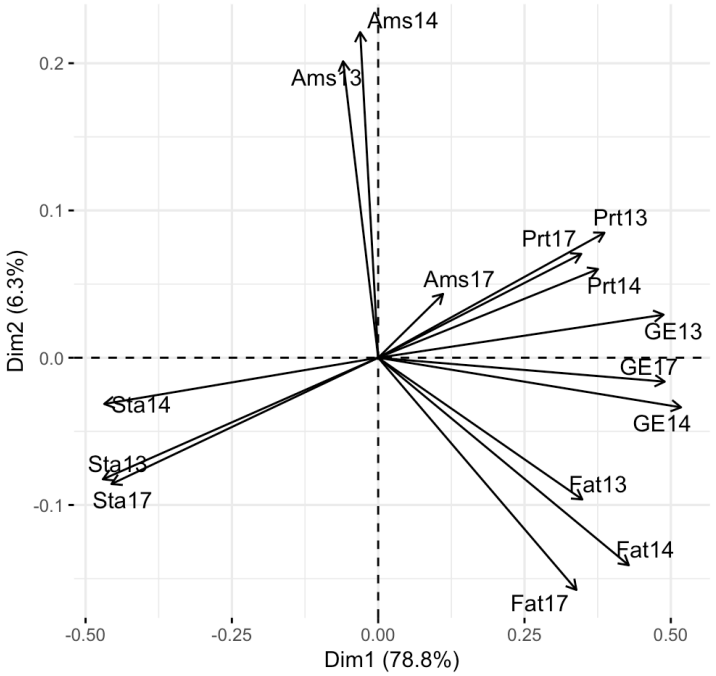


S2 Fig. **Phenotypic distribution of grain composition traits in the GSDP**. Numbers in x-axes represent years. Values are percentage dry basis for protein, fat and starch; gross energy is in Cal/g; and amylose is in percent of starch.

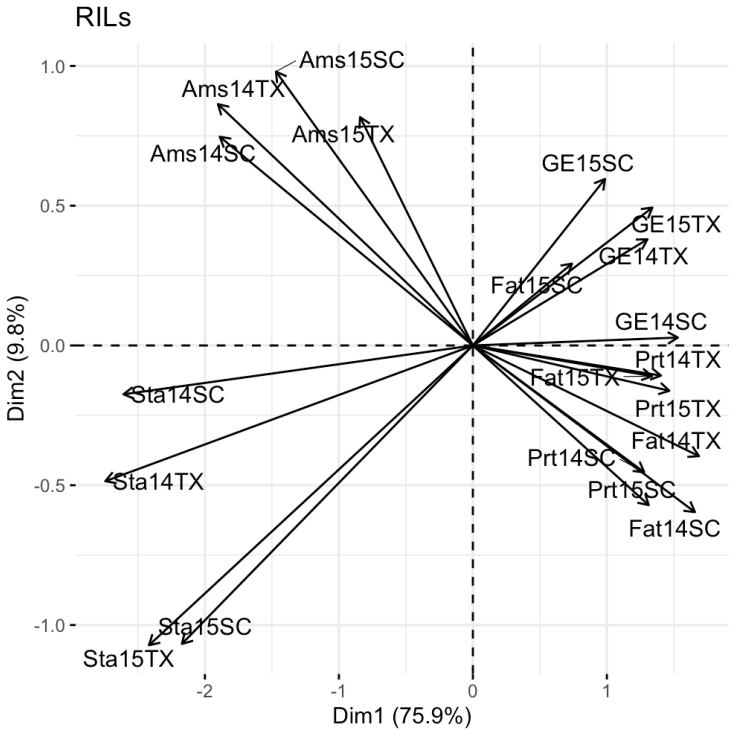


S3 Fig. **PCA analysis of correlation matrix between traits**. **a**. GSDP, and **b**. RILs. Ams: amylose, GE: gross energy, Prt: protein, Sta: starch, SC: South Carolina, TX: Texas. The numbers in the text represent years of the environment.

a.

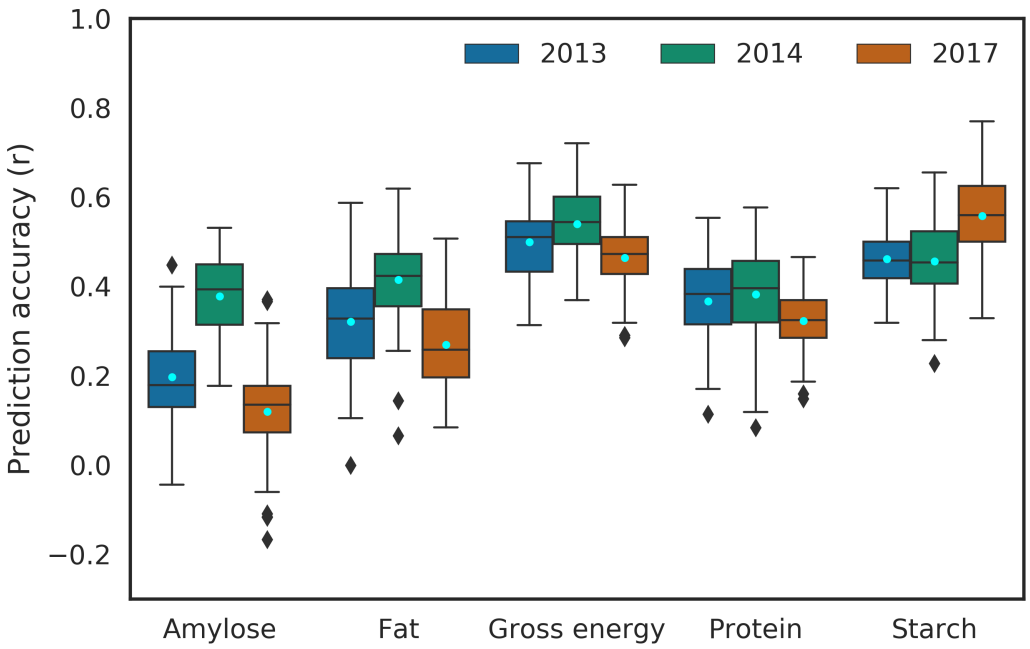


b.

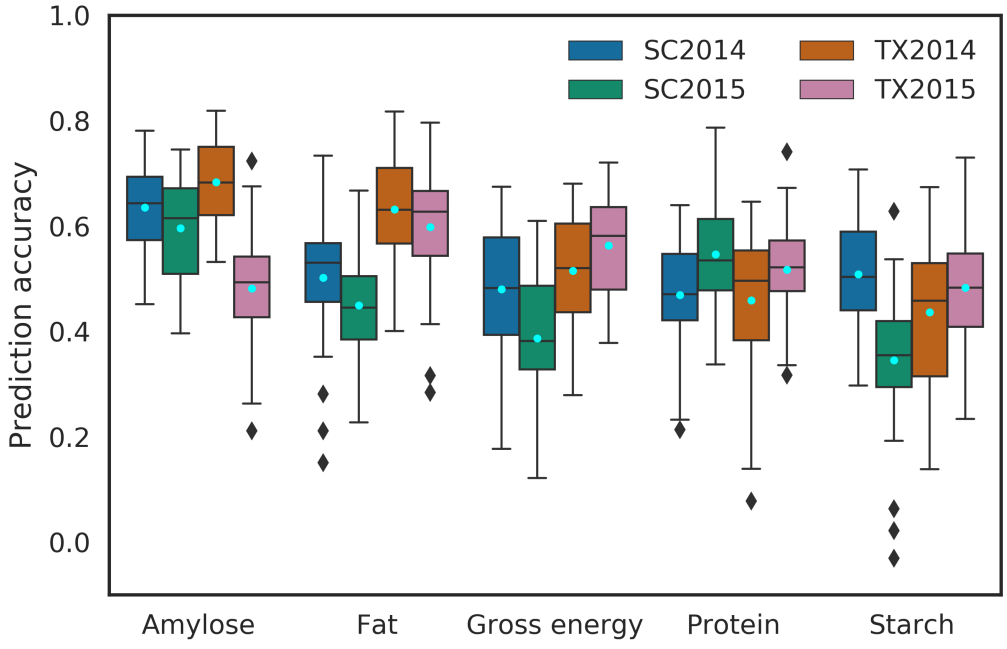


S4 Fig. Prediction accuracy using five-fold CV in Bayesian multi-environment (BME) model. **a**. GSDP, and **b**. RILs. Legend represents the environment/years. SC: South Carolina, TX: Texas. Pale blue dots represent the mean of prediction accuracy.

a.

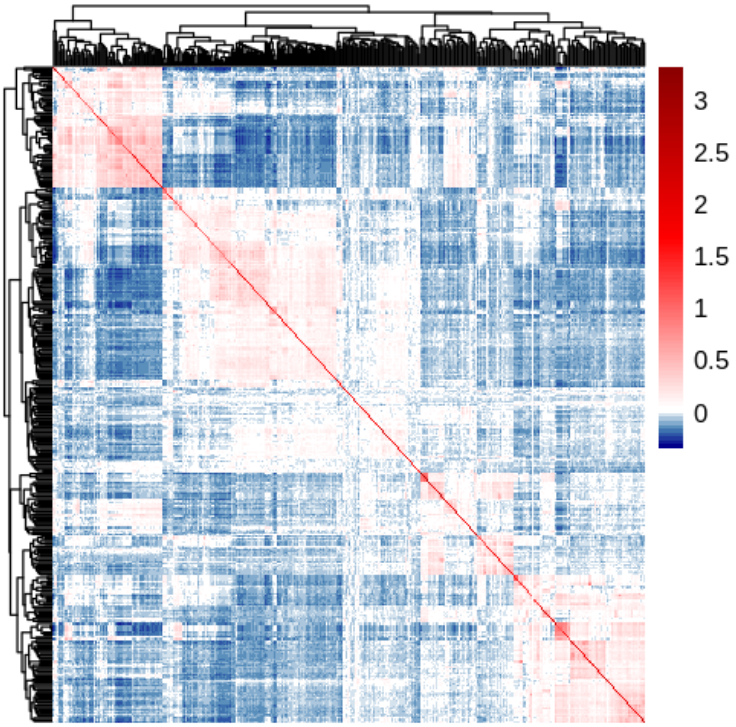


b.

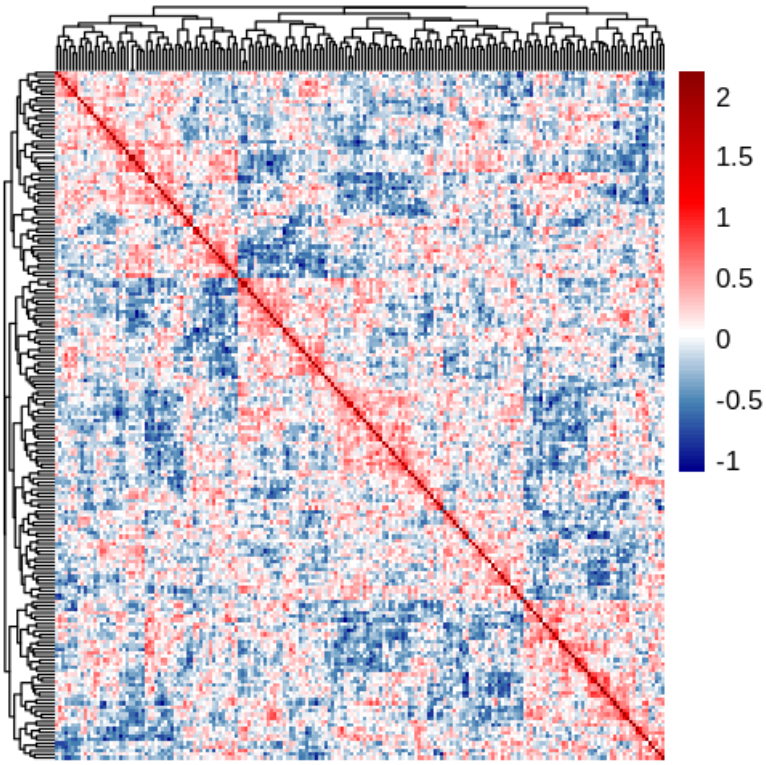


S5 Fig. **Heatmap for genomic relationship matrix calculated using vanRaden (2008)**. **a.** GSDP, **b**. RILs. Trees show hierarchical clustering using Euclidean distance.

a.



b.



S1 Table. **Percent change in prediction accuracy over the single trait single environment model (STSE) model in the GSDP**. BME: Bayesian multi-environment, and BMORS: Bayesian multi-output regressor stacking.

	2013		2014		2017	
Trait	BME	BMORS	BME	BMORS	BME	BMORS
Amylose	-11	66	-5	11	-13	-13
Fat	-24	47	-12	47	-27	58
Gross energy	3	54	-2	40	1	57
Protein	-3	56	-1	55	-8	52
Starch	4	37	-2	38	1	17
Average	-6	52	-4	38	-9	34

S2 Table. **Percent change in prediction accuracy over the single trait single environment model (STSE) model in the RILs**. BME: Bayesian multi-environment, and BMORS: Bayesian multi-output regressor stacking.

Trait	SC2014		SC2015		TX2014		TX2015	
	BME	BMORS	BME	BMORS	BME	BMORS	BME	BMORS
Amylose	2	28	0	28	-1	25	12	43
Fat	5	33	1	15	2	18	2	20
Gross energy	7	28	-3	27	1	18	3	17
Protein	10	51	1	23	5	60	-4	33
Starch	5	36	-4	40	7	54	4	37
Average	6	35	-1	27	3	35	3	30

Appendix C Supplementary File Chapter 4

Genome-wide association and gene network analysis for starch and protein in sorghum

Sirjan Sapkota[1,2,*] , J. Lucas Boatwright[2], Kathleen Jordan[2], Richard Boyles[1,3], and Stephen Kresovich[1,2],

1 Department of Plant and Environmental Sciences, Clemson University, Clemson, SC, USA
2 Advanced Plant Technology Program, Clemson University, Clemson, SC, USA
3 Pee Dee Research and Education Center, Clemson University, Florence, SC, USA

* correspondence: ssapkot@g.clemson.edu

Supplementary File

Supplementary Table S1. Variance components for the linear mixed models fit using genetic and environmental variables.

	Intercept	Geno	Year	Geno × Year	Year × Rep	Residual	Total	H^2
Starch	68.272	3.364	0	1.07	0.141	2.922	7.497	0.799
Protein	12.044	0.869	0.494	0.279	0.066	1.155	2.863	0.753

Supplementary Table S2. Pariwise linkage disequilibrium between association SNPs and their neighboring SNPs. R^2: correlation coefficient, Chr: Chromosome. SNP: single nucleotide polymorphism.

Chr SNP1	Position SNP1	Chr SNP2	Position SNP2	R^2
4	60623675	4	60577500	0.012
4	60623675	4	60623655	0.513
4	60623675	4	60623675	1
4	60623675	4	60624201	0.941
4	60623675	4	60624444	0.921
4	60623675	4	60626903	0.001
4	63400335	4	63380566	0.004
4	63400335	4	63400335	1
4	63400335	4	63426489	0.031
4	64019590	4	64018526	0.009
4	64019590	4	64019122	0.745
4	64019590	4	64019577	1
4	64019590	4	64019590	1
4	64019590	4	64019619	1
4	64019590	4	64028856	0.014
8	51715166	8	51695627	0.057
8	51715166	8	51715166	1
8	51715166	8	51719632	0.219
8	51715166	8	51719659	0.225
8	51715166	8	51719688	0.216
8	51715166	8	51719704	0.993
8	51715166	8	51720767	0.938
8	51715166	8	51721062	0.816
8	51715166	8	51721065	0.794
8	51715166	8	51726098	0.812
8	51715166	8	51727032	0.061

Supplementary Table S3. KEGG Pathway enrichment results for significantly associated genomic regions. FDR: false discovery rate, Observed and Expected refer to the gene count for the network.

Region	Pathway	Observed	Expected	FDR	Genes
Chr-4	Nitrogen metabolism	3	33	1.6E-06	Sb04g024300,Sb04g034470,Sb07g022750
Chr-4	Biosynthesis of unsaturated fatty acids	2	44	0.0004	Sb04g029900,Sb04g029920
Chr-4	Fatty acid metabolism	2	80	0.00085	Sb04g029900,Sb04g029920
Chr-8	Protein processing in endoplasmic reticulum	27	200	1.58E-30	Sb01g005860,Sb01g010460,Sb01g011310,Sb01g013390, Sb01g039390,Sb01g039450,Sb01g039460,Sb01g039470, Sb01g039480,Sb01g039530,Sb01g039780,Sb02g020380, Sb02g021850,Sb02g029650,Sb03g039360,Sb03g041830, Sb04g001140,Sb04g027330,Sb04g030160,Sb06g015020, Sb07g028940,Sb08g003340,Sb08g009580,Sb08g016560, Sb08g018750,Sb09g022580,Sb10g030240
Chr-8	Spliceosome	15	189	5.42E-14	Sb01g011310,Sb01g039390,Sb01g039450,Sb01g039460, Sb01g039470,Sb01g039480,Sb01g039530,Sb01g040250, Sb02g021850,Sb03g039360,Sb03g044450,Sb08g009580, Sb08g015280,Sb08g018750,Sb09g022580
Chr-8	Endocytosis	12	150	1.76E-11	Sb01g011310,Sb01g039390,Sb01g039450,Sb01g039460, Sb01g039470,Sb01g039480,Sb01g039530,Sb02g021850, Sb03g039360,Sb08g009580,Sb08g018750,Sb09g022580
Chr-8	RNA degradation	7	109	1.84E-06	Sb01g017050,Sb01g020010,Sb01g041170,Sb02g028570, Sb04g000370,Sb09g026970,Sb10g001120
Chr-8	Plant-pathogen interaction	6	162	0.00019	Sb03g006570,Sb03g028430,Sb04g021850,Sb07g028940, Sb08g016560,Sb10g030240
Chr-8	Protein export	4	54	0.00022	Sb01g010460,Sb03g041830,Sb04g001140,Sb08g003340